EoS-FM: Can an Ensemble of Specialist Models act as a Generalist Feature Extractor?

Pierre Adorni[1]✉, Minh-Tan Pham[1], Stéphane May[2], Sébastien Lefèvre[1,3]

[1] IRISA, Université Bretagne Sud, UMR 6074, Vannes, France
[2] Centre National d'Études Spatiales (CNES), Toulouse, France
[3] UiT The Arctic University of Norway, Tromsø, Norway
{pierre.adorni,minh-tan.pham,sebastien.lefevre}@irisa.fr,
stephane.may@cnes.fr

Abstract. Recent advances in foundation models have shown great promise in domains such as natural language processing and computer vision, and similar efforts are now emerging in the Earth Observation community. These models aim to generalize across tasks with limited supervision, reducing the need for training separate models for each task. However, current strategies, which largely focus on scaling model size and dataset volume, require prohibitive computational and data resources, limiting accessibility to only a few large institutions. Moreover, this paradigm of ever-larger models stands in stark contrast with the principles of sustainable and environmentally responsible AI, as it leads to immense carbon footprints and resource inefficiency. In this work, we present a novel and efficient alternative: EoS-FM[1], an Ensemble-of-Specialists framework for building Remote Sensing Foundation Models (RSFMs). Our method decomposes the pre-training process into training lightweight, task-specific ConvNeXtV2 specialists that can be frozen and reused. This modular approach offers strong advantages in efficiency, interpretability, and extensibility. Moreover, it naturally supports federated training, pruning, and continuous specialist integration, making it particularly well-suited for collaborative and resource-constrained settings. Our framework sets a new direction for building scalable and efficient RSFMs.

Keywords: Foundation Model · Earth Observation · Remote Sensing · Model Ensembling

1 Introduction

The idea of building a foundation model for remote sensing is an appealing goal to pursue. A single model capable of handling different tasks with far fewer labels than a specialized model would be a huge benefit to the community. In recent years, the Earth Observation (EO) research community has put strong effort into

[1] code and pretrained model are available at https://github.com/pierreadorni/eos-fm.

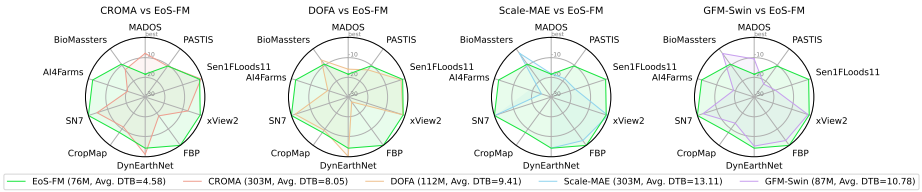

Fig. 1: The proposed EoS-FM demonstrates strong and consistent performance across 11 remote sensing tasks, emerging as the most balanced foundation model among those evaluated on the Pangaea Benchmark [27], despite having fewer parameters. For each method, we show the number of parameters and the average DTB (Distance To Best) metric (lower is better) which will be described in Sec. 4.1.

developing such models, mainly through the use of upscaling techniques [6,10,17]. This approach has already shown success in several areas of Deep Learning and Computer Vision, helping models learn more general and robust features by scaling up both model size and dataset size.

However, while upscaling improves the state of the art, the massive computational cost, energy consumption, and carbon footprint of training multi-billion-parameter models raise serious concerns about accessibility, reproducibility, and environmental impact. Their maintenance and deployment remain out of reach for many academic or operational EO actors, and their size hampers use on edge devices critical for disaster response, agriculture, and climate monitoring.

We argue that the path toward general-purpose Remote Sensing Foundation Models (RSFMs) should emphasize modularity, efficiency, and sustainability over sheer scale. To this end, we introduce a framework to train RSFMs piece by piece, following an Ensemble-of-Specialists (EoS) paradigm. Instead of a single monolithic model, we combine a diverse set of smaller encoders, each specializing in a subset of modalities or tasks, and aggregate their representations. Unlike classical ensembles, which combine full task-specific predictors to improve performance on a single task, our framework combines frozen pretrained encoders as feature extractors to construct a shared representation that generalizes across many downstream tasks. This shifts the role of ensembling from a task-level performance optimization tool to a representation-level mechanism for building general-purpose foundation models. Our main contributions are as follows:

1. A modular and scalable RSFM architecture: We introduce an Ensemble-of-Specialists (EoS) framework that enables combining multiple pre-trained encoders without retraining them jointly.
2. A learned selection mechanism: We propose a differentiable encoder selection layer that scales each encoder's contribution and enables principled train-time pruning. In the full EoS-FM configuration all encoders contribute (k=N); the selection layer is primarily activated in the compact *EoS-FM Small* variant, where it identifies the six most relevant encoders per task, reducing the parameter count to 22M while retaining competitive performance.

3. Strong and balanced performance: Our model achieves the best *average* performance across all 11 Pangaea tasks compared to state-of-the-art RSFMs, while requiring significantly fewer resources to train and deploy. We use the term *generalist feature extractor* specifically to mean a frozen backbone that achieves balanced transfer across diverse RS tasks simultaneously, without task-specific retraining of the encoder.
4. Built-in sustainability and adaptability: The modular design supports federated learning and pruning, making it naturally suited for low-resource or distributed training scenarios.

2 Related Works

Remote Sensing Foundation Models. In Computer Vision, foundation models, typically large-scale Vision Transformers, are pretrained on vast collections of images to learn general-purpose visual features. These models can then serve as frozen backbones with task-specific decoders or be fine-tuned for a given task, requiring far less labeled data than traditional supervised approaches. The idea of building such models for EO, which involves diverse sensors, multiple spectral bands, and temporal dynamics, has gained significant traction in recent years. Over a hundred vision-only foundation models have been released since 2021 [25], showing a clear trend toward larger architectures and pretraining datasets. Most of these efforts rely on self-supervised pretraining, as unlabeled EO imagery is abundant. Despite using datasets smaller than those in general Computer Vision (with DINOv3's SAT493m [39] being a notable large-scale exception), this approach has achieved strong results. We argue that incorporating supervised data, providing a more semantically meaningful training signal, could further reduce the dataset size needed while maintaining high-quality representations. Some recent works follow this direction [3, 50], although they do not explore its potential for smaller and more efficient RSFMs.

Adaptive, Modality-Agnostic RS Models. Several recent works have proposed foundation models that adapt to heterogeneous inputs without requiring strict band alignment. The *Panopticon* model [44] introduces a multi-token adaptive encoder that jointly learns spatial and spectral representations across optical and SAR modalities, achieving flexible per-task adaptation. Similarly, *SMARTIES* [42] learns a spectrum-aware representation space by projecting arbitrary sensor bands into a shared embedding, enabling cross-sensor and cross-mission generalization. The *Copernicus-FM* [46] framework leverages dynamic hypernetworks and aligned Sentinel data to support arbitrary spectral and non-spectral inputs within a single pre-trained backbone. These approaches share the goal of modality-agnostic feature extraction; however, they differ from our work in that they build a unified backbone through joint pre-training or architectural adaptation. In contrast, EoS-FM constructs a general representation by *selecting independently pre-trained encoders and fusing their representations*, leveraging heterogeneity as a feature rather than consolidating it into a single monolithic model.

Model Ensembling. A long-established strategy for improving machine learning performance is model ensembling [11]. In this approach, multiple models, trained independently or with slight variations, are applied to the same input and their predictions are combined, often through majority voting or averaging, to reduce bias and variance. A recent study has shown that ensembles of smaller models can even outperform single large networks in both accuracy and computational efficiency [23]. More recently, feature-level ensembling has emerged as an effective alternative, where intermediate representations from multiple models are fused before classification or decoding [43], improving robustness and leveraging complementary information.

Unlike prior ensembling approaches, which aim to boost performance on a specific downstream task, our objective is to construct a general-purpose representation that transfers across diverse remote sensing tasks. Rather than combining full predictors trained for the same objective, we aggregate frozen pretrained encoders, each specialized on different datasets or modalities, and use their fused features as a shared backbone. In this sense, EoS-FM reframes ensembling from a task-level performance technique to a representation-level mechanism for building modular foundation models.

Mixture of Experts. Increasing the number of parameters generally improves model capacity but also raises inference costs. *Mixture of Experts (MoE)* architectures address this issue by dividing the model into several smaller subnetworks, or experts, and activating only a subset of them for each input [21]. A routing network learns to select which experts to use, allowing the model to maintain high representational power while reducing computational cost [36]. Our approach shares this modular idea: the encoder is composed of multiple disjoint experts (or specialists). However, unlike traditional MoE systems, we train each expert separately on different tasks and only later train the router and feature fusion layers. This results in a sequentially-built EoS rather than a jointly-trained MoE.

3 Proposed Method

3.1 Architecture Overview

Our proposed architecture, EoS-FM (Fig. 2), is an ensemble of ConvNeXtV2-Atto [49] encoders (3.4M parameters each). We chose the ConvNeXtV2 model for its modernity and efficiency; it is known for performing well despite a small number of parameters. However, one might replace it with other backbones without any significant changes to our architecture. Each encoder is trained individually in a supervised manner on a distinct dataset and task; *e.g.* one encoder may learn flood segmentation from SAR imagery, while another learns land use classification from multispectral data. The training procedure is detailed in Sec. 3.5.

3.2 Band Adaptation

Because the encoders in EoS-FM were trained on heterogeneous modalities and channel layouts, we include a *band adaptation* step that maps the available input

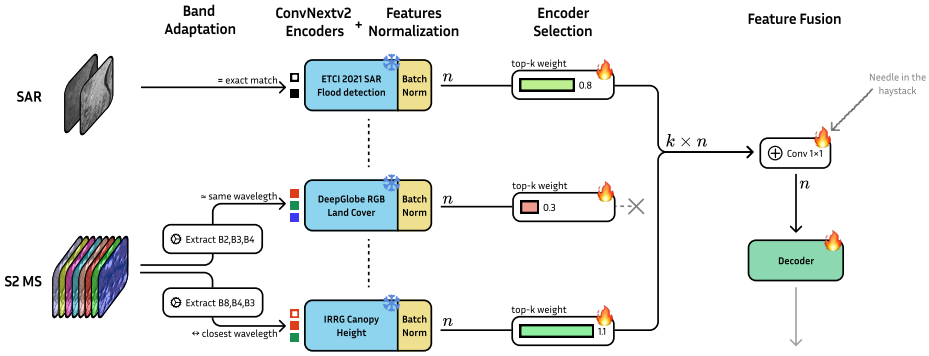

Fig. 2: The EoS-FM backbone first adapts the input bands to each specialist encoder using a spectral-aware mapping, then fuses encoder outputs. A subset of k encoders is selected and aggregated before being passed to the decoder.

bands to each encoder's expected training bands. We use an adaptation layer that computes a single best mapping per encoder using remote-sensing priors.

For each required encoder band, our adaptation layer applies the following priority: (i) exact canonical band-name match, (ii) same center-wavelength match, (iii) nearest-wavelength match within the same modality, and (iv) fallback mapping when no same-modality candidate exists. This design makes adaptation more physically meaningful than purely index-based duplication/subselection and reduces spurious modality mismatches. The adapter also supports simple band expressions (e.g., $VV - VH$) by matching each operand independently and composing the resulting channel on-the-fly.

In practice, using this method, most encoders receive an input that is spectrally very close to what they have seen during training. However, when the modality shift is too large, e.g. when an encoder was pretrained on SAR data but we only have access to optical data for the task at hand, we choose to resort to less physically meaningful adaptations instead of dropping the encoder. The adaptation layer therefore prioritizes physically plausible alignment and falls back to cross-modal adaptation only as a last resort. This design is inspired by SMARTIES [42], which projects heterogeneous sensor bands into a shared spectrum-aware space.

3.3 Encoder Selection

One of the main advantages of our architecture lies in the *modularity of the encoder* and the *specialization* of its subcomponents. Different downstream tasks are likely to rely on distinct types of features, meaning that for a given task, only a subset of the encoders in the ensemble may be truly useful. In other words, the ensemble as a whole performs well across a wide variety of tasks because we can often find a subset of the specialized encoders whose learned representations partially align with the requirements of a new task. If we can identify this subset

Model	HLS	MAD	PAS	S1F	xV2	FBP	DEN	CM	SN7	AI4	BM ↓	Avg DTB ↓
Scale-MAE (303M) [32] ❄	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	62.96	21.47	47.15	13.11
Prithvi (87M) [2] ❄	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	12.20
RemoteCLIP (87M) [24] ❄	76.59	60.00	18.23	74.26	57.41	69.19	31.78	52.05	57.76	25.12	49.79	11.82
SatlasNet (87M) [3] ❄	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	11.64
S12-MAE (22M) [40] ❄	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.80	57.13	24.69	41.07	11.56
SpectralGPT (105M) [20] ❄	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	**36.11**	11.23
S12-Data2Vec (22M) [40] ❄	81.91	44.36	34.32	88.15	51.36	48.82	35.90	54.03	58.23	24.23	41.91	11.20
GFM-Swin (87M) [28] ❄	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	10.97
S12-DINO (22M) [40] ❄	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	10.78
S12-MoCo (22M) [40] ❄	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	10.37
DOFA (112M) [52] ❄	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	9.41
Thor Large (303M) [14]❄	79.47	53.73	39.88	89.55		47.62	37.29	**60.75**	59.71	26.75		9.25
CROMA (303M) [15] ❄	82.42	67.55	32.32	90.89	53.27	51.83	38.29	49.38	59.28	25.65	36.81	8.05
TerraMind Large (303M) [22] ❄	82.93	**75.57**	**43.13**	90.78		63.38	37.89	55.04	59.98	27.47		4.65
EoS-FM (76 M) ❄	87.01	61.96	30.30	89.32	59.79	**70.18**	35.15	52.76	**63.84**	44.1	40.22	**4.58**
EoS-FM Small (22M) ❄	**87.34**	62.93	19.35	90.29	57.56	65.94	32.95	49.14	63.17	38.46	40.20	5.13
UNet (∼8M) [33]🔥	84.51	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	40.39	6.01
ViT-B (87M) [12] 🔥	81.58	48.19	38.53	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	8.78

Table 1: Results on the 11 downstream tasks of the Pangaea benchmark [27]. Our method has the best or second best performance on 5 different datasets, and exhibit the best average performance, on par with TerraMind Large. (DTB = Distance To Best). ↓ Means lower is better.

of relevant encoders, we can prune the others, yielding a much smaller model with negligible loss in performance. The central challenge, then, is to determine *which encoders to keep* for each task.

To address this encoder selection problem, we introduce a lightweight *selection layer* placed immediately after the feature normalization stage, inspired by Mixture-of-Experts (MoE) routing. Let the ensemble contain N encoders $\{E_1, E_2, \ldots, E_N\}$, each producing a feature map $F_i = E_i(x)$ from an input image x. We assign to each encoder a learnable *selection weight* w_i, initialized to 1. Before fusion, each encoder's output is scaled by its corresponding weight:

$$\tilde{F}_i = w_i \, F_i.$$

During **training**, we optimize these weights jointly with the task loss under an explicit sparsity regularization term,

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \, \mathcal{R}(w),$$

where we use a *topk-tail-ℓ_1* penalty. Let $a_i = |w_i|$, and let $a_{(1)} \geq \cdots \geq a_{(N)}$ denote these magnitudes sorted in descending order. The regularizer penalizes only the tail beyond rank k:

$$\mathcal{R}(w) = \sum_{j=k+1}^{N} a_{(j)}.$$

Equivalently, this is the total ℓ_1 mass minus the top-k mass:

$$\mathcal{R}(w) = \sum_{i=1}^{N} |w_i| - \sum_{j=1}^{k} a_{(j)}.$$

This simple objective keeps optimization fully differentiable almost everywhere while directly controlling sparsity through k: reducing k increases the penalized tail mass and encourages only k dominant encoders.

During **validation/inference**, we switch to a hard top-k decision based on the learned (regularized) weights:

$$\tilde{F}_i = \begin{cases} w_i F_i, & \text{if } i \in \text{Top-}k(w), \\ 0, & \text{otherwise.} \end{cases}$$

This train/validation split yields a stable optimization phase and a sparse, interpretable expert subset at evaluation time. This approach is similar in spirit to the experts pruning proposed by MAPEX [18], which also removes modality-specific experts to obtain a lighter model at downstream time. However, while MAPEX prunes expert modules from a single jointly pre-trained Mixture-of-Experts backbone based on modality routing, our method selects among independently pre-trained encoders and performs representation-level selection rather than structural compression of a monolithic model.

Note that in the full EoS-FM configuration, k is set equal to the total number of encoders ($k = N = 22$), meaning all encoders contribute to the representation and no pruning occurs at inference. The selection mechanism is only active when $k < N$: in the compact EoS-FM Small variant, we set $k = 6$, constraining the training-time selection to retain only the six most relevant encoders for each downstream task. This is discussed further in section 4.2.

3.4 Feature Fusion

The selected feature maps are then fused by a single 1×1 convolution layer, which computes linear combinations of the ensemble outputs to reduce dimensionality to a reasonable target size, by default the usual feature size of a single ConvNextv2-Atto encoder.

The fused features are then passed to a decoder to produce the final output. The ensemble functions as a foundation model: for new downstream tasks, all encoders remain frozen, and only the selection, linear fusion and decoder layers are trained. One could argue that the fusion layer is technically part of the encoder, making comparisons with fully frozen foundation models slightly unfair. However, since the fusion layer is minimal in size, we observe similar performance with or without it in practice. Keeping an explicit fusion layer is nevertheless more efficient, as it prevents the decoder from having to process excessively large feature maps.

3.5 Training

We train 22 *ConvNeXtV2-Atto* encoders across 17 datasets and 4 modalities. The datasets were selected from the *torchgeo* library [41] to cover a broad range of input types, including RGB, IR, multispectral, and SAR imagery, as well as

Encoder Name	Dataset	Modality	Bands	Images	Task
eurosat-s2	EuroSat [19]	S2 MS	13	16,200	Scene Cls.
caffe	Caffe [16]	SAR	3	13,090	Calving Fronts Seg.
sen12ms-s1	Sen12MS [35]	SAR	3	130,379	Land Cover Seg.
sen12ms-s2	Sen12MS [35]	S2 MS	13	130,379	Land Cover Seg.
deepglobe-lcc	DeepGlobe LCC [8]	RGB	3	13,000	Land Cover Seg.
bigearthnet-s1	BigEarthNetV2 [7]	S1	3	237,871	Scene Cls.
eurosat	EuroSat [19]	RGB	3	16,200	Scene Cls.
bigearthnet	BigEarthNetV2 [7]	S2 + S1	14	237,871	Scene Cls.
firerisk	Firerisk [38]	RGB	3	70,331	Fire risk Cls.
bigearthnet-rgb	BigEarthNetV2 [7]	RGB	3	237,871	Scene Cls.
sen12ms-rgb	Sen12MS [35]	RGB	3	130,379	Land Cover Seg.
imagenet	ImageNet [9]	RGB	3	1,281,167	FCMAE
minifrance	MiniFrance [5]	RGB	3	472,476	Land Cover Seg.
etci2021	ETCI2021	SAR	3	21,600	Flood Seg.
opencanopy	OpenCanopy [13]	IR-R-G	3	66,368	Canopy Height Reg.
potsdam	Potsdam [34]	IR-R-G	3	1,200	Land Cover Seg.
cloud_cover	Cloud Cover [30]	IR-R-G	3	11,748	Cloud Seg.
inria_aerial	Inria [26]	RGB	3	15,500	Building Seg.
landcoverai	LandCover.ai [4]	RGB	3	7,470	Land Cover Seg.
levircdplus	Levir CD+ [37]	RGB	3	510	Building CD
loveda	LoveDA [45]	RGB	3	2,522	Land Cover Seg.
kenyacroptype	CV4A Kenya Crop Type [29]	S2 MS	13	4,836	MultiTemporal Seg.

Table 2: Summary of the datasets used to train the EoS-FM ensemble. Notations S1: Sentinel-1; S2 MS: Sentinel-2 Multispectral, SAR: Synthetic Aperture Radar, IR-R-G: Infrared-Red-Green, Cls: Classification, Seg: Segmentation, CD: Change Detection.

different tasks such as classification, segmentation, and change detection. The trained encoders and related data modalities are described in Table 2.

Some datasets provide multiple modalities for the same geographic samples. For instance, BigEarthNet includes both Sentinel-1 and Sentinel-2 data. In such cases, we train multiple encoders by selecting subsets of the available bands or modalities, forcing each encoder to specialize in the information contained in specific inputs. For example, we train three encoders on BigEarthNet: one using both Sentinel-1 and Sentinel-2 data (14 channels total), one using only Sentinel-1 only and one using the RGB channels extracted from the Sentinel-2 image only. In total, EoS-FM is trained on 1,085,101 unique samples. Some samples are viewed multiple times under different modalities due to the modality subsampling strategy, as previously illustrated with BigEarthNet.

Each encoder is initialized from a *ConvNeXtV2-Atto* model pretrained in a self-supervised manner on ImageNet (via the `timm` library [48]), and fine-tuned until convergence on its respective dataset. When input images are too large, we use a tiling strategy with a convenient patch size (*e.g.*, 512×512 for MiniFrance).

4 Experiments

4.1 Downstream Tasks

We follow the evaluation protocol of the Pangaea Benchmark [27], where the EoS-FM encoder ensemble is frozen and a UperNet decoder [51] is fine-tuned for 80 epochs with a batch size of 8. The best checkpoint is selected based on validation mIoU, and final results are reported using test mIoU for segmentation

Model	HLS	MAD	PAS	S1F	xV2	FBP	DEN	CM	SN7	AI4	BM ↓	Avg DTB ↓
RemoteCLIP (87M) ❄	69.4	20.57	17.19	62.22	53.75	56.23	34.43	19.86	43.11	23.85	53.32	15.95
Scale-MAE (303M) ❄	75.47	21.47	22.86	64.74	**56.06**	48.75	35.27	13.44	49.68	26.66	54.16	14.77
GFM-Swin (87M) ❄	67.23	28.19	21.47	62.57	53.45	55.58	28.16	27.21	39.48	32.88	49.3	14.17
SatlasNet (87M) ❄	74.79	29.87	16.76	83.92	44.07	37.86	34.64	29.08	49.78	13.91	44.38	13.86
S12-Data2Vec (22M) ❄	74.38	17.86	33.09	81.91	41.6	37.27	33.63	34.11	40.66	22.85	46.52	13.81
S12-MAE (22M) ❄	76.6	18.44	31.06	84.81	39.84	35.56	30.59	35.29	40.51	23.6	43.76	13.65
Prithvi (87M) ❄	77.73	21.24	33.56	86.28	35.08	29.98	32.28	27.71	36.78	35.04	41.19	13.48
SpectralGPT (105M) ❄	71.82	20.29	34.53	83.12	35.81	39.51	35.33	31.06	36.31	37.35	**39.44**	12.46
DOFA (112M) ❄	71.98	23.77	27.68	82.84	55.6	27.82	**39.15**	29.91	46.1	27.74	46.03	12.38
S12-MoCo (22M) ❄	73.11	19.47	32.51	79.58	41.15	35.57	32.24	36.54	49.46	37.97	44.83	11.82
S12-DINO (22M) ❄	75.93	23.47	36.62	84.95	41.02	34.63	32.78	38.44	41.15	37.91	42.74	10.78
CROMA (303M) ❄	76.44	32.44	32.8	87.22	46.54	37.39	36.08	36.77	42.15	38.48	40.25	8.79
Terramind Base (86M) ❄	77.39	44.06	**39.96**	84.43		54.00	37.35	35.65	43.21	38.59		5.72
Thor Base (86M) ❄	76.90	40.67	38.93	86.29		42.80	35.21	**42.23**	55.94	38.90		5.36
EoS-FM (76 M) ❄	81.69	44.13	31.94	82.24	53.92	**66.73**	**39.79**	23.61	**61.32**	38.03	41.05	**3.67**
EoS-FM Small (22M) ❄	**82.97**	**45.38**	15.18	86.45	51.42	62.59	34.84	20.16	52.38	37.46	51.67	7.78
UNet (∼8M) 🔥	79.46	24.3	29.53	**88.55**	46.77	52.58	35.59	13.88	46.08	34.84	40.39	10.14
ViT-B (87M) 🔥	75.92	10.18	38.44	81.85	44.85	56.53	35.39	27.76	36.01	**39.20**	44.89	11.05

Table 3: Results on the 11 downstream tasks of the Pangaea benchmark [27], using only **10% of the training data**. Our method has the best performance on five different datasets, and exhibit the best average performance. ↓ means lower is better.

tasks and test RMSE for the single regression task, BioMassters. We compare against models for which results on Pangaea have already been published in a peer-reviewed paper[2]. All performance metrics in Tab. 1 and Tab. 3 are reported.

The Pangaea benchmark assesses the generalization ability of foundation models, how transferable their learned features are across diverse Earth observation tasks, by counting how often each model ranks in the top two positions across benchmarks. While this metric highlights models achieving state-of-the-art (SOTA) performance on multiple datasets, it overlooks models that consistently perform close to the SOTA across all tasks. We argue that the latter quality, balanced generalization, is essential for a robust, well-rounded foundation model.

To better quantify this, we employ the *Average Distance To Best* (Avg. DTB) metric proposed in [1], which computes the mean absolute difference between a model's performance and the best score achieved on each dataset. Unlike *"number of top-2s"* metric reported in the Pangaea Benchmark, Avg. DTB rewards models that perform reliably well on all tasks rather than excelling at a few while failing on others.

We compare our method to all models included in the original release of the Pangaea benchmark, as well as to two recent baselines that evaluated themselves on Pangaea in their respective works: Terramind [22] and Thor [14]. The authors of Terramind and Thor did not report results on two datasets from the Pangaea benchmark (xView2 and BioMassters), following the recommendation of the Pangaea authors regarding reproducibility considerations. In our case, we were able to obtain stable results on these two datasets and therefore report

[2] As of the submission deadline, March 5, 2026.

the performance of EoS-FM across all tasks. The DTB metric is computed using only the available performance values for each model: over 11 tasks for all models except Terramind and Thor, for which it is computed over 9 tasks.

As shown in Table 1, EoS-FM achieves the lowest Avg. DTB (4.58) among all tested models, on par with TerraMind Large (4.65). This indicates that, despite having one-fourth the parameters of the largest tested models, EoS-FM is the most consistent across all 11 downstream tasks. TerraMind and EoS-FM are the only tested models that outperform the supervised baselines (UNet Avg. DTB 6.01). EoS-FM ranks first on the FiveBillionPixels, HLS Burn Scars, and Spacenet 7 Change Detection dataset, and second on AI4SmallFarms and xView2, demonstrating its ability to generalize across very heterogeneous remote sensing modalities: from high-resolution optical imagery to coarse multi-temporal data. On the AI4Farms dataset in particular, it significantly surpasses all other RSFMs, being the only frozen model to approach the performance of a fully trained encoder.

Another notable finding concerns the UNet baseline. The original Pangaea authors observed that, in the full data regime, this fully supervised baseline outperformed all RSFMs on average, suggesting a gap between pretraining and task-specific adaptation. Using our new metric, this finding stands: the best frozen model of the original Pangaea benchmark (CROMA) achieves an average DTB of 8.05, which underperforms the 6.01 of Unet. However, both TerraMind and our model close this gap completely: they significantly surpasses the UNet in Avg. DTB (4.58 and 4.65 vs. 6.01), while maintaining the benefits of frozen encoder adaptation and pretraining versatility.

In contrast, several other foundation models exhibit strong but uneven performance. For example, CROMA excels on MADOS and Sen1Floods11 but lags on FiveBillionPixels and AI4Farms. Scale-MAE achieves SOTA scores on xView2 and SpaceNet7, but remains the less performing one in terms of Avg. DTB (13.11). SpectralGPT and the S12 family of models also show solid results on individual benchmarks but fall short in overall consistency. These fluctuations underline that many RSFMs remain specialized, not truly general-purpose: a limitation EoS-FM directly addresses through its encoder ensemble design.

Label Scarcity. One of the central promises of foundation models is the ability to perform well with limited supervision. In Earth Observation, where labeled data is expensive and often imbalanced, particularly for rare events such as natural disasters, this ability is crucial. Following the Pangaea protocol, we train a UperNet decoder on only 10% of the available labels per task, using stratified sampling to preserve class balance.

Results in Table 3 confirm that EoS-FM retains its advantage under severe label scarcity. It achieves the best performance on five datasets and maintains the best average DTB (3.67), outperforming all baselines. This stability under low-data regimes highlights that EoS-FM's features are both rich and transferable, requiring fewer labeled examples to adapt effectively. Notably, EoS-FM and Thor are the only models whose Avg. DTB improves under label scarcity, reflecting

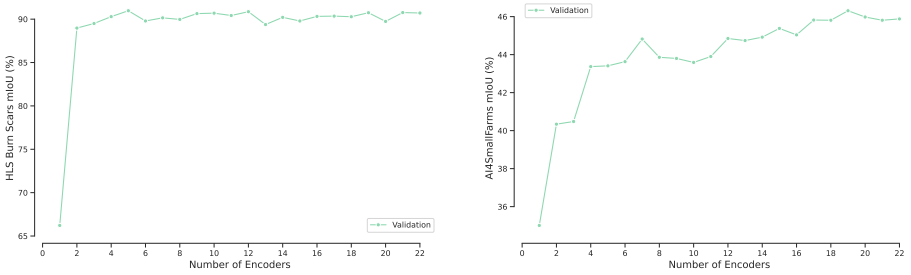

Fig. 3: Ablation study: increasing the number of encoders increases the performance of the ensemble in a frozen setting. The AI4SmallFarms dataset benefits more from a larger ensemble than the HLS Burn Scars does. The validation mIoU is reported.

smaller performance degradation rather than higher absolute scores. The smaller variant, EoS-FM Small (22M parameters), even outperforms the full ensemble on several datasets (HLS Burn Scars, MADOS, Sen1Flood11), despite its lower overall Avg. DTB (7): when labels are scarce, reducing the number of feature maps can mitigate overfitting, yielding a cleaner representation space and better generalization.

Overall, these results collectively demonstrate that EoS-FM is not only well-rounded and consistent but also highly label-efficient, establishing it as a strong candidate for a general-purpose RSFM.

4.2 Scaling & Pruning

The results presented in the previous section were obtained with a fixed version of EoS-FM composed of 22 encoders. This configuration was chosen arbitrarily, based on our available resources and our intuition about what constitutes a diverse training set. However, the modular design of our architecture makes it straightforward to extend.

To analyze how downstream performance scales with the ensemble size, we conduct an ablation study varying the number of encoders k from 1 to 22. Increasing k effectively enlarges the subset of encoders selected for feature fusion until the maximum ensemble size is reached. This experiment is performed on the HLS Burn Scars and AI4SmallFarms datasets, with results presented in Fig. 3. Two key observations emerge:

- **Performance scales with ensemble size.** for HLS Burn Scars, mIoU rises quickly from 66% with 1 encoder to 90% with 4 encoders, and for AI4Small farms, mIoU increases more slowly from 35% with 1 encoder to 45% with 17 encoders. This experiment confirms that leveraging more diverse encoders improves overall performance.
- **Performance stagnates after a few encoders.** On HLS Burn Scars, after the first 4 encoders, performance stagnates around 90%, with slight variations due to the stochastic nature of this single-run experiment. This

suggests that only a fraction of the encoders are actually able to contribute to some tasks, and reinforces the usefulness of our proposed train-time pruning strategy, described below.

The top-k selection layer can serve as a mechanism for *train-time pruning*, effectively controlling the model's size during fine-tuning on a new dataset. Given a large ensemble of encoders, this layer enables the construction of task-specific lightweight variants of EoS-FM by retaining only the k most relevant encoders while training the decoder. To illustrate this capability, we benchmark EoS-FM on the Pangaea datasets using $k = 6$, which reduces the total parameter count to 22M—comparable to a ResNet-50. This compact configuration, referred to as *EoS-FM Small*, achieves competitive results while maintaining the efficiency benefits of a much smaller model. The corresponding performances are reported in Table 1 and Table 3.

Analysis of the selected encoders. After training the EoS-FM Small variant, we analyzed which encoders were retained or discarded by the selection layer. In many cases, the selected subsets appear intuitively relevant in a post-hoc analysis. For instance, on Sen1Floods11 — a dataset combining Sentinel-1 and Sentinel-2 imagery — the selected encoders include eurosat (S2), bigearthnet (S1+S2), sen12ms (S2), caffe (S1), etci2021 (S1), and bigearthnet (S1), all of which were pretrained on compatible modalities.

Similarly, for Five Billion Pixels, a very high-resolution (VHR) dataset, the selected encoders predominantly correspond to high-resolution pretraining sources, including eurosat (S2), landcover.ai (VHR), ImageNet, Potsdam (VHR), open-canopy (VHR), and loveda (VHR).

However, we also observe several intriguing behaviors. The encoders are not selected with equal frequency across tasks; some appear consistently more useful within the Pangaea Benchmark. Moreover, certain selections are less immediately intuitive. For example, on SpaceNet7 Change Detection: a high-resolution building construction dataset, the Inria Aerial encoder (HR building segmentation) was not selected, despite apparent task similarity, whereas the sen12ms (S1) encoder, pretrained exclusively on SAR imagery, was retained.

These observations suggest that encoder utility is not determined solely by superficial task similarity or modality alignment. A deeper investigation of these dynamics is further explored in the supplementary material.

4.3 Feature normalization

We motivate the use of Batch Normalization over the encoder features by drawing from previous work on feature distillation [31, 47], which highlights the importance of normalizing teacher features to mitigate issues caused by differences in magnitude. However, the cited approaches implement normalization differently: [31] introduces a custom invertible normalization scheme to recover the original feature space, while [47] employs Layer Normalization. Since our

Dataset	∅	LayerNorm	BatchNorm
HLS Burn Scars	80.52	84.34 +3.82	**84.41** +3.89
MADOS	**66.06**	64.05 -2.05	62.42 -3.64
Sen1FLoods11	87.95	**89.22** +1.27	89.02 +1.07
SpaceNet 7 CD	54.85	55.43 +0.58	54.76 -0.07
AI4SmallFarms	37.55	38.04 +0.49	39.56 +2.01
Avg.	65.39	66.22 +0.83	**66.03** +0.64

Table 4: Effect of different feature normalization strategies on ensemble performance. val mIoU is reported for each dataset.

Model	GFLOPs	FPS
SSL4EO (ResNet50)	8.2	**2664**
EoS-FM Small	**7.2**	664
DOFA	17.8	568
EoS-FM	25.4	208
CROMA	126.4	200
Scale-MAE	120.0	168

Table 5: Computational efficiency comparison on RTX 5000 GPU (bs=8, 224×224).

pipeline does not require inverting the transformation, we opted for a simpler, non-parametric Batch Normalization layer.

To empirically validate this choice, we conducted an ablation study comparing no normalization, Layer Norm, and Batch Norm across multiple datasets (Table 4). On average, adding a feature normalization step does improve performance, while the difference between the two tested normalization schemes is marginal.

4.4 Computational Efficiency

In addition to predictive performance, we evaluate the computational efficiency of EoS-FM and compare it with representative Remote Sensing Foundation Models (RSFMs). Table 5 reports the computational cost in GFLOPs and inference throughput in frames per second (FPS). All measurements are conducted on an NVIDIA RTX 5000 Ada GPU with batch size 8 and input resolution 224×224. GFLOPs are computed as $2\times$GMACs using the `ptflops` library, and FPS is measured during inference with frozen encoders and a UperNet decoder.

The results show that EoS-FM achieves competitive efficiency despite combining multiple pretrained encoders. The lightweight variant, EoS-FM Small, requires only 7.2 GFLOPs, making it more efficient than DOFA while maintaining strong downstream performance (Sec. 4.1). The full EoS-FM model requires significantly fewer FLOPs than large-scale foundation models such as CROMA and Scale-MAE, reducing computational cost by approximately $5\times$ while achieving superior balanced generalization.

We note that inference throughput does not scale linearly with FLOPs, and the two metrics should be interpreted as complementary rather than equivalent. FLOPs measure theoretical arithmetic cost; FPS reflects wall-clock throughput, which additionally depends on memory bandwidth, kernel fusion, and hardware utilization. Architectures such as ResNets and Transformers benefit from mature, highly-optimized CUDA kernels that give them a throughput advantage disproportionate to their FLOP count. In contrast, EoS-FM runs multiple independent ConvNeXtV2 encoders, incurring scheduling and memory overhead that depresses FPS relative to the raw FLOP comparison. EoS-FM Small largely

avoids this penalty by activating only 6 encoders, reaching 664 FPS at 7.2 GFLOPs.

5 Discussion

Generalization through specialization. Our results show that EoS-FM achieves strong and balanced performance across diverse tasks, suggesting an alternative path to general-purpose representations. While most foundation models rely on large-scale joint pretraining of a single backbone, EoS-FM aggregates multiple pretrained specialists. Despite being optimized for specific tasks or modalities, their fusion yields representations that generalize consistently. This indicates that general-purpose features may emerge not only from scaling individual models, but also from composing diverse specialized encoders.

We acknowledge that some pretraining datasets partially overlap in task type with the Pangaea evaluation suite (e.g., SpaceNet7 and xView2 relate to LevirCD+ and Inria Aerial; ETCI2021 relates to Sen1Floods11). Disentangling genuine representation-level generalization from broad specialist coverage is therefore non-trivial. Several lines of evidence suggest, however, that EoS-FM does more than retrieve the nearest specialist: (i) performance scales monotonically up to 4–5 encoders (Fig. 3), so no single specialist dominates; (ii) several strong Pangaea tasks (HLSBurnScars, MADOS, PASTIS, Sen1Floods11) have no close counterpart in the pretraining pool; and (iii) manually curated modality-aligned subsets consistently underperform automatically selected ones (Supp. Sec. D.1), suggesting the fusion exploits complementary information beyond simple modality matching. We nonetheless acknowledge this as a genuine limitation and encourage future work to disentangle the two effects under more controlled conditions.

Standalone feature extraction. Like other RSFMs, EoS-FM generalizes to new objectives by attaching a lightweight decoder to frozen encoders and finetuning only the decoder and fusion layer. However, the raw concatenation of all encoder features produces prohibitively large representations, limiting direct use as a universal embedding model. Future work could explore training a compact fusion module as a shared projection head to enable efficient standalone feature extraction.

Federated learning. The modular design naturally supports federated learning. Encoders can be trained independently on local or proprietary datasets and later integrated into the ensemble without sharing raw data. This enables collaborative and privacy-preserving scaling of Earth Observation models, aligning with sustainability goals.

Broader impact. Although introduced for remote sensing, the Ensemble-of-Specialists paradigm is domain-agnostic. Composing lightweight, task-specialized encoders and fusing their representations may offer a modular alternative to increasingly large monolithic models in any multimodal or heterogeneous setting.

Relationship to multi-task pretraining. An alternative direction would be to train a single backbone jointly on all 22 tasks. While conceptually appealing,

this approach faces well-known challenges including conflicting gradients, imbalanced task importance, and increasingly complex loss weighting as task diversity grows. The ensemble approach sidesteps these difficulties by training each specialist independently. We view joint multi-task pretraining as a complementary direction and leave a direct comparison to future work.

6 Conclusion

We have introduced EoS-FM, an ensemble-based framework for building efficient and modular foundation models in remote sensing. By training lightweight specialized encoders on diverse datasets and fusing their representations, our method achieves competitive performance compared to much larger models, while remaining scalable and easily prunable. These results highlight the potential of composing FMs from smaller, domain-specialized components rather than relying solely on monolithic architectures.

Beyond performance, our approach introduces several methodological contributions. We propose a model design that is inherently modular and scalable, allowing new encoders to be added incrementally and outdated ones to be replaced without retraining the entire ensemble. Our top-k selection layer enables train-time pruning, making it possible to derive compact task-specific variants such as *EoS-FM Small* from a larger ensemble. Finally, the architecture naturally supports federated learning setups, where encoders can be trained independently in distributed data environments and later integrated through the fusion module. Together, these elements illustrate a different path toward sustainable and collaborative RSFM development, grounded in flexibility, efficiency, and extensibility.

Acknowledgments

This project was funded in part by the Université Bretagne Sud, the Centre National d'Etudes Spatiales (CNES), and the Agence Nationale de la Recherche (ANR) through two projects: DREAMS (ANR-25-CE56-4384-01) and Cluster SequoIA (ANR-23-IACL-0009). It was granted HPC computing and storage resources by GENCI-IDRIS under grant 2025-AD011015819R1 on the V100 and H100 partitions of the Jean Zay supercomputer. The authors would like to thank the NASA Earth Science Data Systems Program, NASA Digital Transformation AI/ML thrust, and IEEE GRSS for organizing the ETCI competition.

References

1. Adorni, P., Pham, M.T., May, S., Lefèvre, S.: Towards efficient benchmarking of foundation models in remote sensing: A capabilities encoding approach. In: Proceedings of the Computer Vision and Pattern Recognition Conference Workshops. pp. 3096–3106 (2025) 9

2. et al., J.J.: Foundation Models for Generalist Geospatial Artificial Intelligence (arXiv:2310.18660) (Nov 2023) 6
3. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: Satlaspretrain: A large-scale dataset for remote sensing image understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16772–16782 (2023) 3, 6
4. Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T., Zambrzycka, A.: Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1102–1110 (2021) 8
5. Castillo-Navarro, J., Le Saux, B., Boulch, A., Audebert, N., Lefèvre, S.: Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. Machine Learning **111**(9), 3125–3160 (2022) 8
6. Cha, K., Seo, J., Lee, T.: A billion-scale foundation model for remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing pp. 1–17 (2024). https://doi.org/10.1109/JSTARS.2024.3401772 2
7. Clasen, K.N., Hackel, L., Burgert, T., Sumbul, G., Demir, B., Markl, V.: reBEN: Refined bigearthnet dataset for remote sensing image analysis. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2025) 8
8. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 172–181 (2018) 8
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 8
10. Dias, P., Tsaris, A., Bowman, J., Potnis, A., Arndt, J., Yang, H.L., Lunga, D.: Oreole-fm: successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems. pp. 597–600 (2024) 2
11. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000) 4
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (Oct 2020) 6
13. Fogel, F., Perron, Y., Besic, N., Saint-André, L., Pellissier-Tanon, A., Schwartz, M., Boudras, T., Fayad, I., d'Aspremont, A., Landrieu, L., et al.: Open-canopy: Towards very high resolution forest monitoring. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1395–1406 (2025) 8
14. Forgaard, T., Reksten, J.H., Waldeland, A.U., Marsocci, V., Longépé, N., Kampffmeyer, M., Salberg, A.B.: Thor: A versatile foundation model for earth observation climate and society applications. arXiv preprint arXiv:2601.16011 (2026) 6, 9
15. Fuller, A., Millard, K., Green, J.R.: CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. In: Thirty-Seventh Conference on Neural Information Processing Systems (Nov 2023) 6

16. Gourmelon, N., Seehaus, T., Braun, M., Maier, A., Christlein, V.: Calving fronts and where to find them: A benchmark dataset and methodology for automatic glacier calving front extraction from sar imagery. Earth System Science Data Discussions **2022**, 1–37 (2022) 8
17. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., Li, Y.: SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27662–27673. IEEE, Seattle, WA, USA (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.02613 2
18. Hanna, J., Scheibenreif, L., Borth, D.: Mapex: Modality-aware pruning of experts for remote sensing foundation models. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–11 (2026) 7
19. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019) 8
20. Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J.: Spectralgpt: Spectral remote sensing foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5227–5244 (2024). https://doi.org/10.1109/TPAMI.2024.3362475 6
21. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991) 4
22. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al.: Terramind: Large-scale generative multimodality for earth observation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7383–7394 (2025) 6, 9
23. Kondratyuk, D., Tan, M., Brown, M., Gong, B.: When ensembling smaller models is more efficient than single large models. In: Proceedings of the Computer Vision and Pattern Recognition Conference Workshops (2020) 4
24. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024). https://doi.org/10.1109/TGRS.2024.3390838 6
25. Lu, S., Guo, J., Zimmer-Dauphinee, J.R., Nieusma, J.M., Wang, X., Wernke, S.A., Huo, Y., et al.: Vision foundation models in remote sensing: A survey. IEEE Geoscience and Remote Sensing Magazine (2025) 3
26. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: 2017 IEEE International geoscience and remote sensing symposium (IGARSS). pp. 3226–3229. IEEE (2017) 8
27. Marsocci, V., Jia, Y., Le Bellier, G., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., et al.: Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. IEEE geoscience and remote sensing magazine (2025) 2, 6, 8, 9
28. Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C.: Towards Geospatial Foundation Models via Continual Pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16806–16816 (2023) 6
29. Radiant Earth Foundation: CV4A Competition Kenya Crop Type Dataset (2020) 8

30. Radiant Earth Foundation: Sentinel-2 cloud cover segmentation dataset (version 1) (2022). https://doi.org/10.34911/rdnt.hfq6m7, [Accessed: 2025-11-11] 8
31. Ranzinger, M., Barker, J., Heinrich, G., Molchanov, P., Catanzaro, B., Tao, A.: Phi-s: Distribution balancing for label-free multi-teacher distillation (2024), https://arxiv.org/abs/2410.01680, [Accessed 2026-06-30] 12
32. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023) 6
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 6
34. Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U.: The isprs benchmark on urban object classification and 3d building reconstruction (2012) 8
35. Schmitt, M., Hughes, L., Qiu, C., Zhu, X.: Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **IV-2/W7**, 153–160 (09 2019). https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019 8
36. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR) (2017), https://openreview.net/forum?id=B1ckMDqlg, [Accessed: 2026-06-30] 4
37. Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B.: S2looking: A satellite side-looking dataset for building change detection. Remote Sensing **13**(24), 5094 (2021) 8
38. Shen, S., Seneviratne, S., Wanyan, X., Kirley, M.: Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning. In: 2023 international conference on digital image computing: techniques and applications (DICTA). pp. 189–196. IEEE (2023) 8
39. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 3
40. Stewart, A., Lehmann, N., Corley, I., Wang, Y., Chang, Y.C., Ait Ali Braham, N.A., Sehgal, S., Robinson, C., Banerjee, A.: Ssl4eo-l: Datasets and foundation models for landsat imagery. Advances in Neural Information Processing Systems **36**, 59787–59807 (2023) 6
41. Stewart, A.J., Robinson, C., Corley, I.A., Ortiz, A., Lavista Ferres, J.M., Banerjee, A.: TorchGeo: Deep learning with geospatial data. ACM Transactions on Spatial Algorithms and Systems (Dec 2024). https://doi.org/10.1145/3707459 7
42. Sumbul, G., Xu, C., Dalsasso, E., Tuia, D.: Smarties: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5569–5578 (2025) 3, 5
43. Ullah, Z., Kim, J.: Hierarchical deep feature fusion and ensemble learning for enhanced brain tumor mri classification. Mathematics **13**(17), 2787 (2025) 4
44. Waldmann, L., Shah, A., Wang, Y., Lehmann, N., Stewart, A., Xiong, Z., Zhu, X.X., Bauer, S., Chuang, J.: Panopticon: Advancing any-sensor foundation models for earth observation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2204–2214 (2025) 3

45. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran Associates, Inc. (2021) 8
46. Wang, Y., Xiong, Z., Liu, C., Stewart, A.J., Dujardin, T., Bountos, N.I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., et al.: Towards a unified copernicus foundation model for earth vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9888–9899 (2025) 3
47. Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., Guo, B.: Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint arXiv:2205.14141 (2022) 12
48. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019), [Accessed 2026-06-30] 8
49. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023) 4
50. Wu, K., Zhang, Y., Ru, L., Dang, B., Lao, J., Yu, L., Luo, J., Zhu, Z., Sun, Y., Zhang, J., et al.: A semantic-enhanced multi-modal remote sensing foundation model for earth observation. Nature Machine Intelligence pp. 1–15 (2025) 3
51. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018) 8
52. Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L., Camps-Valls, G., Zhu, X.X.: Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation (arXiv:2403.15356) (Jun 2024). https://doi.org/10.48550/arXiv.2403.15356 6