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

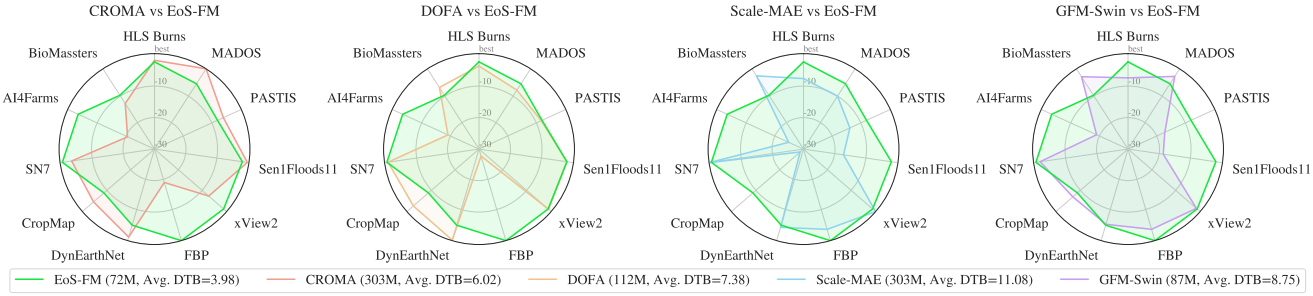

Figure 1. The proposed EoS-FM demonstrates strong and consistent performance across 11 remote sensing tasks, emerging as the most balanced foundation model among those evaluated on the Pangaea Benchmark [25], despite having fewer parameters. For each method, we show the number of parameters and the average DTB (Distance To Best) metric (lower is better) which will be described in Sec. 4.1.

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: an Ensemble-of-Specialists framework for building Remote Sensing Foundation Models (RSFMs). Our method decomposes the training process into 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. All codes and pretrained models are available at https://github.com/pierreadorni/EoS-FM.

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 developing such models, mainly through the use of upscaling techniques [5, 9, 17]. This approach has already shown success in several areas of Deep Learning and Computer Vision, helping models learn

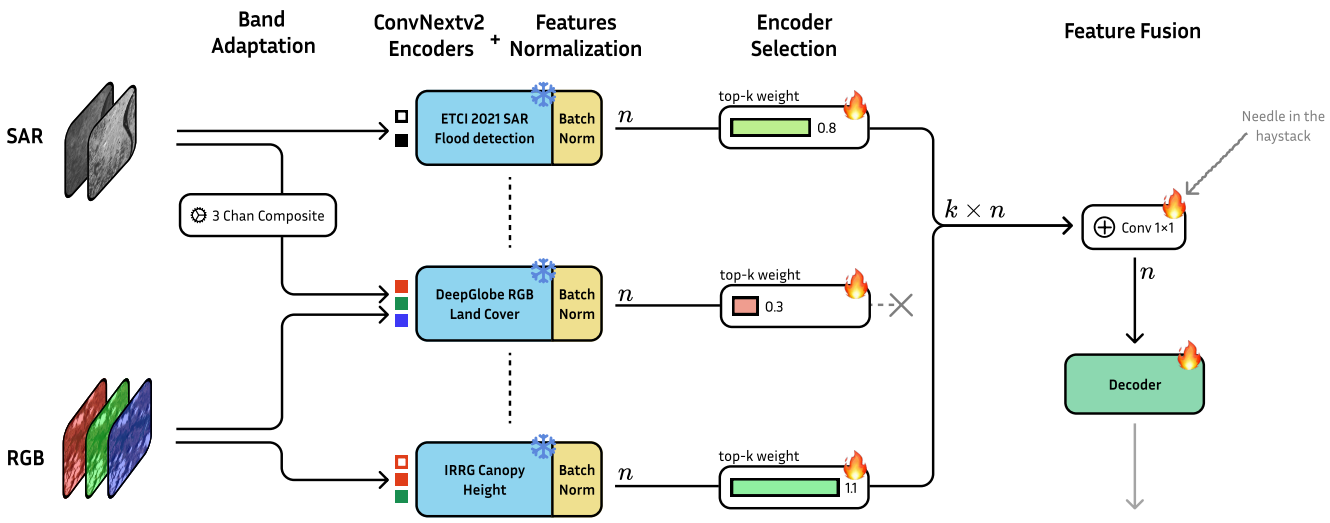

Figure 2. The EoS-FM Backbone adapts any given input to a multitude of formats using band duplication and selection to extract as many feature maps as possible, and then fuse them. Each encoder produces n feature maps; a subset of k encoders is then selected for fusion, and their $k \times n$ feature maps are aggregated into n fused feature maps before being passed to the decoder.

more general and robust features by scaling up both model size and dataset size.

However, while upscaling improves the state of the art, it is not an efficient or sustainable strategy. The massive computational resources and energy consumption required to train these models raise serious concerns about their accessibility, reproducibility, and environmental impact. The carbon footprint associated with training multi-billion-parameter models can rival that of entire research institutions, and their maintenance and fine-tuning remain out of reach for many academic or operational EO actors. Moreover, their size hampers deployment on edge devices and low-resource settings, which are critical for real-world applications such as 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. 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 dynamically selects the most relevant subset of encoders for a given input or task.
3. Strong performance and efficiency: Our model achieves competitive performance compared to state-of-the-art RSFMs on the Pangaea Benchmark [25], while requiring significantly fewer resources to train and deploy.
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 [23], 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 [35] 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

Encoder Name	Dataset	Modality	# Bands	# Images	Task
eurosat-s2	EuroSat [18]	S2 MS	13	16,200	Scene Cls.
caffe	Caffe [16]	SAR	3	13,090	Calving Fronts Seg.
sen12ms-s1	Sen12MS [31]	SAR	3	130,379	Land Cover Seg.
sen12ms-s2	Sen12MS [31]	S2 MS	13	130,379	Land Cover Seg.
deepglobe-lcc	DeepGlobe LCC [7]	RGB	3	13,000	Land Cover Seg.
bigearthnet-s1	BigEarthNetV2 [6]	S1	3	237,871	Scene Cls.
eurosat	EuroSat [18]	RGB	3	16,200	Scene Cls.
bigearthnet	BigEarthNetV2 [6]	S2 + S1	14	237,871	Scene Cls.
firerisk	Firerisk [34]	RGB	3	70,331	Fire risk Cls.
bigearthnet-rgb	BigEarthNetV2 [6]	RGB	3	237,871	Scene Cls.
sen12ms-rgb	Sen12MS [31]	RGB	3	130,379	Land Cover Seg.
imagenet	ImageNet [8]	RGB	3	1,281,167	FCMAE Reconstruction
minifrance	MiniFrance [4]	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 [30]	IR-R-G	3	1,200	Land Cover Seg.
cloud_cover	Cloud Cover [14]	IR-R-G	3	11,748	Cloud Seg.
inria_aerial	Inria [24]	RGB	3	15,500	Building Seg.
landcoverai	LandCover.ai [3]	RGB	3	7,470	Land Cover Seg.
levircdplus	Levir CD+ [33]	RGB	3	510	Building CD
loveda	LoveDA [39]	RGB	3	2,522	Land Cover Seg.

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

this direction [2, 43], although they do not explore its potential for smaller and more efficient RSFMs.

Moddel Ensembling. A long-established strategy for improving machine learning performance is model ensembling [10]. In this approach, multiple models, trained independently or with slight variations, are applied to the same input. Their predictions are combined, often through majority voting or averaging, to reduce individual model 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 [21]. More recently, feature-level ensembling has emerged as an effective alternative, where the intermediate representations (rather than the final predictions) from multiple models are fused before classification or decoding [38]. This approach improves representation robustness and leverages complementary information from diverse models. In this work, we follow this principle by combining the representations extracted by several specialized encoders, each acting as a domain expert, before passing the fused features to a shared decoder. This ensemble-of-specialists design allows us to harness the diversity of specialized models while maintaining a unified and efficient downstream architecture.

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 sub-networks, or experts, and activating only a subset of them for each input [20]. A routing network learns to select which experts to use, allowing the model to maintain high representational power while reducing computational cost [32]. 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 [42] 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. 3.6.

3.2. Band Adaptation

Because the encoders operate on inputs with different numbers of spectral bands, we include a *band adaptation* step before feeding the inputs to the encoders. This step applies predefined rules that map available bands to the required ones through band duplication and subset selection. A rule is defined by a tuple (*available bands*, *required bands*) and a list of indices. For instance, our implementation contains the following mappings from SAR Sentinel-1 (S1) and Optical Sentinel-2 (S2) imagery to RGB space:

1. $(B_{S2}, B_{RGB}) \rightarrow (B_{S2})_{4,3,2}$
2. $(B_{S2}, B_{RGB}) \rightarrow (B_{S2})_{8,4,3}$
3. $(B_{S1}, B_{RGB}) \rightarrow (B_{S1})_{0,1,1}$

Rules 1 and 2 extract the R–G–B (B4-B3-B2) and IR–R–G (B8-B4-B3) bands respectively from S2 data, while rule 3 creates a pseudo-RGB image from SAR data by duplicating the VH channel: (VV, VH, VH). When multiple rules share the same condition (like rules 1 and 2), we apply all of them and extract features for each resulting input. This increases the number of output feature maps: even a small ensemble of encoders can produce many feature maps, one for each (*encoder*, *band adaptation rule*) pair.

A natural question arises: *why feed SAR data into an encoder pretrained on RGB images?* Our reasoning lies in the feature fusion mechanism. While an encoder trained on modality A is unlikely to extract semantically meaningful features from modality B, it may still capture low-level visual structures (such as edges or textures) that remain useful after fusion [46]. We hypothesize that such partial transferability can be beneficial, especially when combined with feature selection or fusion layers. We therefore extract as many feature maps as possible and rely on the fusion layer to perform the *needle-in-a-haystack* selection of relevant information.

3.3. Feature Map Normalization

The encoders composing our ensemble were trained on different tasks with different decoders, and therefore operate in distinct embedding spaces, producing feature maps that follow different distributions. Additionally, feeding an encoder data from a distribution it has never encountered, like feeding RGB images to a SAR-pretrained encoder, is likely to result in abnormally distributed feature maps. While this is usually not a major issue when the decoder is fully trained, in our case, where the features of multiple encoders are fused, it can cause the fusion layer to rely disproportionately on the *loudest* encoders rather than treating all encoders equally. Figure 3 illustrates this imbalance by showing the variance of feature maps across some encoders in our ensemble.

Previous work on multi-teacher distillation has addressed this issue by normalizing the feature outputs of different teachers to make their magnitudes comparable [27, 40]. Following this idea, we apply a non-parametric batch normalization layer, without scaling or bias, to each encoder's feature map before fusion.

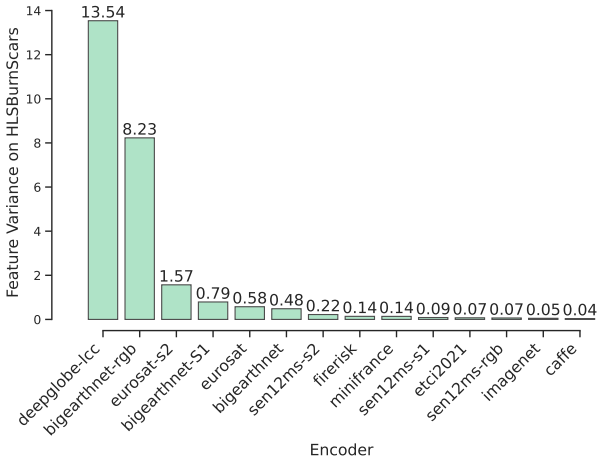

Figure 3. Variance of the feature maps computed by different encoders on the HLS Burn Scars training set. The variance changes a lot between encoders, which could create problems when training an ensemble.

3.4. 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 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.

Model	HLS Burns	MADOS	PASTIS	Sen1Fl	xView2	FBP	DynEarthNet	CropMap	SN7	AI4Farms	BioMass ↓	Avg DTB ↓
Scale-MAE (303M) [28] ❄	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	**62.96**	21.47	47.15	11.08
Prithvi (87M) [12] ❄	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	10.17
RemoteCLIP (87M) [22] ❄	76.59	60.00	18.23	74.26	57.41	69.19	31.78	52.05	57.76	25.12	49.79	9.79
SatlasNet (87M) [2] ❄	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	9.53
S12-MAE (22M) [36] ❄	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.80	57.13	24.69	41.07	9.61
SpectralGPT (105M) [19] ❄	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	**36.11**	9.20
S12-Data2Vec (22M) [36] ❄	81.91	44.36	34.32	88.15	51.36	48.82	35.90	**54.03**	58.23	24.23	41.91	9.17
GFM-Swin (87M) [26] ❄	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	8.75
S12-DINO (22M) [36] ❄	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	8.94
S12-MoCo (22M) [36] ❄	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	8.34
DOFA (112M) [45] ❄	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	7.38
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	6.02
EoS-FM (72 M) ❄	81.92	61.96	30.30	89.32	59.47	**70.89**	34.39	45.06*	62.22	42.71	39.89	**3.81**
EoS-FM Small (22M) ❄	75.93	60.78	27.45	87.02	58.52	68.42	33.19	36.61*	58.95	38.73	44.02	7.29
UNet (∼8M) [29] 🔥	**84.51**	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	40.39	3.98
ViT-B (87M) [11] 🔥	81.58	48.19	**38.53**	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	6.75

Table 2. Results on the 11 downstream tasks of the Pangaea benchmark [25]. Our method has the best performance on FiveBillionPixels (FBP), the second best performance on AI4SmallFarms and Spacenet7 (SN7) Change Detection, and exhibit the best average performance, even surpassing the supervised baselines (DTB = Distance To Best). ↓ means lower is better. *averaged over 3 runs due to significant run variability.

Before fusion, each encoder's output is scaled by its corresponding weight:

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

If a sparsity hyperparameter $k < N$ is defined, we retain only the k encoders with the highest weights $\{w_i\}$, setting all other feature maps to zero:

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

These scaled feature maps are then passed to the fusion layer. Since the same encoder may be reused for multiple input bands or modalities, its weight w_i is consistently applied across all derived feature maps. The multiplicative scaling ensures that gradients can propagate through the selection mechanism, allowing the model to learn the optimal encoder subset jointly with the downstream objective. By default, k is set to the ensemble size, so all feature maps are preserved.

3.5. 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.

3.6. Training

We train 21 *ConvNeXtV2-Atto* encoders across 16 datasets and 4 modalities. The datasets were selected from the *torchgeo* library [37] to cover a broad range of input types, including RGB, IR, multispectral, and SAR imagery, as well as different tasks such as classification, segmentation, and change detection. The trained encoders and related data modalities are described in Table 1.

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 1,080,265 unique samples. Some samples are viewed multiple times under different modalities due to the modality subsampling strategy, as previously illustrated with BigEarthNet.

Model	HLS Burns	MADOS	PASTIS	Sen1Fl	xView2	FBP	DynEarthNet	CropMap	SN7	AI4Farms	BioMass ↓	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	14.27
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	13.09
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	12.48
SatlasNet (87M) ❄	74.79	29.87	16.76	83.92	44.07	37.86	34.64	29.08	49.78	13.91	44.38	12.17
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	12.13
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	11.97
Prithvi (87M) ❄	77.73	21.24	33.56	86.28	35.08	29.98	32.28	27.71	36.78	35.04	41.19	11.79
SpectralGPT (105M) ❄	71.82	20.29	34.53	83.12	35.81	39.51	35.33	31.06	36.31	37.35	**39.44**	10.78
DOFA (112M) ❄	71.98	23.77	27.68	82.84	55.6	27.82	**39.15**	29.91	46.1	27.74	46.03	10.70
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	10.13
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	9.10
CROMA (303M) ❄	76.44	32.44	32.8	87.22	46.54	37.39	36.08	36.77	42.15	38.48	40.25	7.11
EoS-FM (72 M) ❄	71.82	**47.05**	29.24	79.48	55.27	**64.18**	32.05	22.97*	**52.13**	**40.2**	41.82	**4.70**
EoS-FM Small (22M) ❄	71.48	45.53	26.41	80.16	54.06	62.80	32.59	21.83*	51.82	39.88	47.11	5.89
UNet (∼8M) 🔥	**79.46**	24.3	29.53	**88.55**	46.77	52.58	35.59	13.88	46.08	34.84	40.39	8.46
ViT-B (87M) 🔥	75.92	10.18	**38.44**	81.85	44.85	56.53	35.39	27.76	36.01	39.2	44.89	9.36

Table 3. Results on the 11 downstream tasks of the Pangaea benchmark [25], using only **10% of the training data**. Our method has the best performance on four different datasets, and exhibit the best average performance, even surpassing supervised baselines on average. ↓ means lower is better. *averaged over 3 runs due to significant run variability.

Each encoder is initialized from a *ConvNeXtV2-Atto* model pretrained in a self-supervised manner on ImageNet (via the `timm` library [41]), 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 [25], where the EoS-FM encoder ensemble is frozen and a UperNet decoder [44] 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 tasks and test RMSE for the single regression task, BioMassters.

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.

As shown in Table 2, EoS-FM achieves the lowest Avg. DTB (3.81) among all tested models, including much larger ones like CROMA (303M parameters). This indicates that, despite having less than one-fourth the parameters of the largest tested models, EoS-FM is the most consistent across all 11 downstream tasks.

EoS-FM ranks first on the FiveBillionPixels benchmark and second on both AI4Farms and SpaceNet7 Change Detection, 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 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 Pangaea benchmark (CROMA) achieves an average DTB of 6.02, which significantly underperforms the 3.98 of Unet. However, our model closes this gap completely: EoS-FM matches and even slightly surpasses the UNet in Avg. DTB (3.81 vs. 3.98), 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 agricultural datasets like AI4Farms. Scale-MAE achieves two best

scores on xView2 and SpaceNet7, but remains the less performing one in terms of Avg. DTB (11.08). 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 four datasets and maintains the best average DTB (4.70), comfortably outperforming all RSFMs and even the fully supervised UNet (8.46). This stability under low-data regimes highlights that EoS-FM's features are both rich and transferable, requiring fewer labeled examples to adapt effectively.

Interestingly, while some foundation models such as DOFA or Scale-MAE lose approximately 2 points of Avg. DTB when reducing data availability, EoS-FM degrades minimally, from 3.81 to 4.70. This resilience demonstrates that the model's ensemble-based representation is inherently data-efficient. Conversely, the UNet baseline, which was competitive in the full-data regime, loses much more performance under label scarcity, underscoring the typical brittleness of fully supervised methods.

Finally, the smaller variant, EoS-FM Small (22M parameters), provides an additional insight: despite being only one-third the size of the full model, it still performs competitively—achieving an Avg. DTB of 7.29 in the full-data setting and 5.86 under label scarcity, rivaling models with three to four times more parameters. This result confirms that the ensemble mechanism scales down gracefully and can produce lightweight, task-specific models while retaining task-specific performance.

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 21 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.

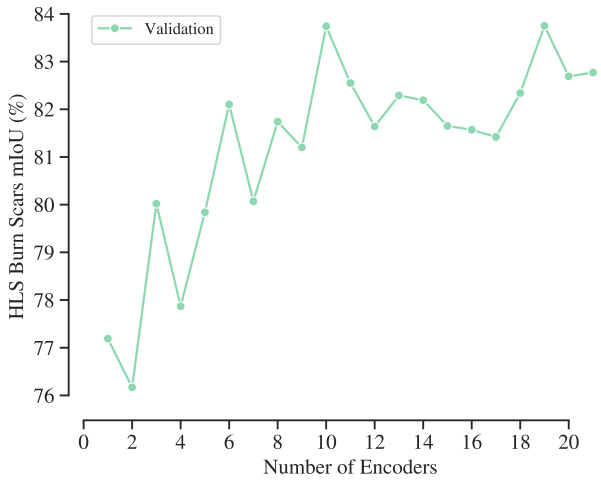

Figure 4. Ablation study: increasing the number of encoders increases the performance of the ensemble in a frozen setting. Experiments are performed on the HLS Burn Scars dataset. The validation mIoU is reported.

To analyze how downstream performance scales with the ensemble size, we conduct an ablation study varying the number of encoders k from 1 to 21. 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 dataset, with results presented in Fig. 4. Two key observations emerge:

- **Performance scales with ensemble size.** mIoU rises from 76% with 2 encoders to 84% with 19 encoders, confirming that leveraging more diverse encoders improves overall performance.
- **Non-monotonic behavior.** Performance gains are not strictly increasing—adding an encoder does not always yield improvement. This volatility is especially pronounced for small values of k, where the choice of encoders strongly impacts results. This suggests that the encoder selection layer may not consistently identify the optimal subset, pointing to potential room for refinement in the selection mechanism.

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 2 and Table 3.

4.3. Feature normalization

Dataset	∅	Layer Norm	Batch Norm
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
CropTypeMapping	15.14	45.13 +29.99	**60.70** +45.56
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.	57.01	62.70 +5.69	**65.15** +8.14

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

We motivate the use of Batch Normalization over the encoder features by drawing from previous work on feature distillation [27, 40], which highlights the importance of normalizing teacher features to mitigate issues caused by differences in magnitude. However, the cited approaches implement normalization differently: [27] introduces a custom invertible normalization scheme to recover the original feature space, while [40] employs Layer Normalization. Since our 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, Batch Normalization yields the best results, largely due to a substantial improvement on the CropTypeMapping dataset (+45.56 mIoU). Interestingly, the MADOS dataset shows the opposite trend: performance improves when normalization is removed altogether, suggesting that the effect of normalization may depend on the underlying feature distribution of each dataset.

5. Discussion

Standalone feature extraction. We proposed a new architecture, EoS-FM, which demonstrates strong and versatile representation capabilities across a wide range of downstream tasks. Similar to other Remote Sensing Foundation Models, it can generalize to new objectives by attaching a lightweight decoder to the frozen encoder and fine-tuning only the decoder — and in our case, the feature fusion layer. However, unlike most RSFMs, EoS-FM is not easily usable as a standalone feature extractor: the raw concatenation of features from all 21 encoders results in prohibitively large feature maps, making direct use of the un-fused representations impractical. Future work could investigate whether a compact, pre-trained fusion layer could act as a general-purpose projection head, enabling the use of EoS-FM as a universal embedding model for remote sensing data.

Federated learning. A key advantage of the proposed architecture lies in its modularity. Each encoder in the ensemble operates independently and contributes to the shared representation space only through the fusion layer. This design naturally lends itself to federated learning: new institutions could locally train additional encoders on their proprietary or geographically specific data, and later integrate them into the global ensemble without sharing the raw data. Such an approach could enable the creation of a continually improving, privacy-preserving foundation model that grows collaboratively across organizations and sensors — an attractive direction for scaling Earth Observation models sustainably.

Broader Impact. Although EoS-FM is presented in the context of remote sensing, its architecture is not tied to this domain. The idea of assembling a model from lightweight, task-specialized encoders and fusing their intermediate representations is broadly applicable wherever data come in multiple modalities or require heterogeneous expertise. Remote sensing is a natural fit because of its combination of optical, multispectral and SAR products, but many other fields face similar challenges. The Ensemble-of-Specialists approach offers a new way to integrate diverse sources of information without resorting to increasingly large monolithic models, opening the door to more efficient and modular foundation models beyond Earth Observation.

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 provided with computing HPC and storage resources by GENCI at IDRIS thanks to the grant 2025-AD011015819R1 on the supercomputer Jean Zay's V100 and H100 partitions. 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] Pierre Adorni, Minh-Tan Pham, Stéphane May, and Sébastien Lefèvre. Towards efficient benchmarking of foundation models in remote sensing: A capabilities encoding approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3096–3106, 2025. 6

[2] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 3, 5

[3] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. 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*, pages 1102–1110, 2021. 3

[4] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9):3125–3160, 2022. 3

[5] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A Billion-scale Foundation Model for Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–17, 2024. 1

[6] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*, 2024. 3

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[9] Philipe Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, H Lexie Yang, and Dalton Lunga. 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*, pages 597–600, 2024. 1

[10] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020. 5

[12] Johannes Jakubik et al. Foundation Models for Generalist Geospatial Artificial Intelligence. (arXiv:2310.18660), 2023. 5

[13] Fajwel Fogel, Yohann Perron, Nikola Besic, Laurent Saint-André, Agnès Pellissier-Tanon, Martin Schwartz, Thomas Boudras, Ibrahim Fayad, Alexandre d'Aspremont, Loic Landrieu, et al. Open-canopy: Towards very high resolution forest monitoring. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1395–1406, 2025. 3

[14] Radiant Earth Foundation. Sentinel-2 cloud cover segmentation dataset (version 1), 2022. [Accessed: 2025-11-11]. 3

[15] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 5

[16] Nora Gourmelon, Thorsten Seehaus, Matthias Braun, Andreas Maier, and Vincent Christlein. 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. 3

[17] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. 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)*, pages 27662–27673, Seattle, WA, USA, 2024. IEEE. 1

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. 3

[19] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. 5

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3

[21] Dan Kondratyuk, Mingxing Tan, Matthew Brown, and Boqing Gong. When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570*, 2020. 3

[22] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 5

[23] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Steven A Wernke, Yuankai Huo, et al. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025. 2

[24] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 3

[25] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. (arXiv:2412.04204), 2024. 1, 2, 5, 6

[26] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 5

[27] Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *URL https://arxiv. org/abs/2410.01680*, 2024. 4, 8

[28] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 5

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[30] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. 2012. 3

[31] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 3

[32] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 3

[33] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. 3

[34] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. 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)*, pages 189–196. IEEE, 2023. 3

[35] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[36] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36:59787–59807, 2023. 5

[37] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. TorchGeo: Deep learning with geospatial data. *ACM Transactions on Spatial Algorithms and Systems*, 2024. 5

[38] Zahid Ullah and Jihie Kim. Hierarchical deep feature fusion and ensemble learning for enhanced brain tumor mri classification. *Mathematics*, 13(17):2787, 2025. 3

[39] J Wang, Z Zheng, A Ma, X Lu, and Y Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 3

[40] Y Wei, H Hu, Z Xie, Z Zhang, Y Cao, J Bao, D Chen, and B Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 4, 8

[41] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 6

[42] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 3

[43] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, et al. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, pages 1–15, 2025. 3

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[45] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. (arXiv:2403.15356), 2024. 5

[46] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 4

Supplementary Material — 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

1. Introduction

Our main paper introduced a new framework for building Remote Sensing Foundation Models (RSFMs), along with a pre-trained model, EoS-FM. We described the architecture, training, evaluation, as well as some experiments. This supplementary document provides additional experiment and analyses supporting the findings presented in the main paper. We expand on the stability of our fine-tuning results on the Pangaea benchmark, provide full numerical tables for multi-run evaluation, and discuss methodological choices that were only briefly introduced in the main text. These additions are meant to give a more complete view of EoS-FM's behavior across datasets, clarify aspects of the experimental setup, and ensure full reproducibility of our approach.

2. Stability of Pangaea benchmark fine-tunings

In the main paper we reported a single run per downstream dataset, except for *CropTypeMapping*, for which we provided the mean of three runs. To assess the stability of EoS-FM fine-tuning on the Pangaea benchmark, we later performed two additional independent runs per dataset using identical hyperparameters and independently sampled random seeds. Tab. 1 reports the resulting metrics, along with their mean and standard deviation. Due to time constraints, we were unable to perform additional fine-tunings for *BioMassters*, and thus it is omitted from the table.

Dataset	Run 1	Run 2	Run 3	Mean $\pm$ Std
HLS Burn Scars	81.92	79.83	80.45	80.73 ± 1.07
MADOS	61.96	60.22	61.79	61.32 ± 0.96
PASTIS	30.30	29.34	28.77	29.47 ± 0.77
Sen1Floods11	89.32	89.41	88.31	89.01 ± 0.61
xView2	59.47	58.36	58.76	58.86 ± 0.56
FiveBillionPixels	70.89	68.56	69.07	69.51 ± 1.22
DynEarthNet	34.39	35.29	36.36	35.35 ± 0.99
CropTypeMapping	49.45^*	39.88^*	45.89^*	$45.07^* \pm 4.84$
SpaceNet 7	62.22	62.64	60.44	61.77 ± 1.17
AI4SmallFarms	42.71	38.61	39.02	40.11 ± 2.26

Table 1. Stability study: three independent fine-tunings per downstream dataset. All metrics are mIoU. Values marked with * correspond to the original three-run average from the main paper for *CropTypeMapping*, for which additional runs were not repeated.

Notes on methodology. All runs follow exactly the same training procedure used in the main experiments: identical optimizer, learning rate schedule, number of epochs, data augmentations, and dataset splits. The only source of variation is the

Model	HLS Burns	MADOS	PASTIS	Sen1Fl	xView2	FBP	DynEN	CropMap	SN7	AI4Farms	BioMass ↓	Avg DTB ↓
Scale-MAE (303M) [9] ❄	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	**62.96**	21.47	47.15	10.96
Prithvi (87M) [3] ❄	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	10.05
RemoteCLIP (87M) [6] ❄	76.59	60.00	18.23	74.26	57.41	69.19	31.78	52.05	57.76	25.12	49.79	9.67
S12-MAE (22M) [11] ❄	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.80	57.13	24.69	41.07	9.48
SatlasNet (87M) [1] ❄	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	9.41
SpectralGPT (105M) [5] ❄	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	**36.11**	9.08
S12-Data2Vec (22M) [11] ❄	81.91	44.36	34.32	88.15	51.36	48.82	35.90	**54.03**	58.23	24.23	41.91	9.05
S12-DINO (22M) [11] ❄	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	8.82
GFM-Swin (87M) [8] ❄	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	8.62
S12-MoCo (22M) [11] ❄	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	8.22
DOFA (112M) [12] ❄	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	7.25
CROMA (303M) [4] ❄	82.42	**67.55**	32.32	90.89	53.27	51.83	38.29	49.38	59.28	25.65	36.81	5.90
EoS-FM (72 M) ❄	80.73	61.32	29.47	89.01	58.86	**69.51**	35.35	45.06	61.77	40.11	39.89	4.33
UNet (∼8M) [10] 🔥	**84.51**	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	40.39	**3.85**
ViT-B (87M) [2] 🔥	81.58	48.19	**38.53**	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	6.63

Table 2. Results on the 11 downstream tasks of the Pangaea benchmark [7], averaged over three runs (except *BioMassters*). EoS-FM achieves the best performance on *FiveBillionPixels*, ranks second on *AI4SmallFarms*, and obtains the strongest Avg. DTB among frozen models, approaching the performance of a fully supervised UNet. Lower is better for the ↓ metrics (DTB = Distance To Best).

random seed, which affects data ordering and other nondeterministic operations. For each dataset, *Run 1* corresponds to the value reported in the main paper, except for *CropTypeMapping*, where we had already conducted and reported the mean of three runs.

As expected, the results show strong stability across runs, with most standard deviations close to 1. Two exceptions stand out: *CropTypeMapping*, with a standard deviation of 4.84, and *AI4SmallFarms*, with 2.26. While the instability of the former was already known from the main experiments, the latter was not observed previously.

Effects on benchmark performance. Tab. 2 summarizes the updated Pangaea Benchmark results, where all metrics are averaged over three runs (except for *BioMassters*, as noted above). These more precise values do not change our overall findings: EoS-FM remains the top-performing frozen model on *Five Billion Pixels*, ranks second on *AI4SmallFarms* (just behind the supervised UNet), and maintains the best Avg. DTB among all frozen approaches. Its performance is slightly below that of the supervised UNet, whereas our initial single-run experiments placed both methods at roughly similar levels. We attribute this slight shift to the small but non-negligible variability inherent in fine-tuning on the Pangaea benchmark.

References

[1] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020. 2

[3] Johannes Jakubik et al. Foundation Models for Generalist Geospatial Artificial Intelligence. (arXiv:2310.18660), 2023. 2

[4] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 2

[5] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. 2

[6] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 2

[7] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. (arXiv:2412.04204), 2024. 2

[8] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 2

[9] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 2

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[11] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36:59787–59807, 2023. 2

[12] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. (arXiv:2403.15356), 2024. 2