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, UMR 6074, Université Bretagne Sud, 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 —largely centered on scaling the number of parameters and dataset size— require prohibitive computational and data resources, limiting accessibility to only a few large institutions. 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.

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 [8, 10, 9]. 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, it is not a very efficient method. Training huge models requires an enormous number of images, which makes it nearly impossible for most players to do. Their large size also creates problems for inference, especially when deploying on edge devices. We believe that efficiency should be a key concern from the very start when designing foundation models. To address this, we propose a framework to train a Remote Sensing Foundation Model (RSFM) piece by piece, using an Ensemble-Of-Specialists (EoS) approach. This design naturally supports federated learning and pruning, which we see as essential features for building efficient and lightweight RSFMs.

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, 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 [13] 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 [14, 6], although they do not explore its potential for smaller and more efficient RSFMs.

Model Ensembling. A long-established method for improving machine learning performance is *model ensembling* [2]. 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. The underlying assumption is that individual models make different errors, so combining their outputs can reduce the overall error rate. In this work, we adopt a related idea, but instead of merging the *predictions* of multiple models on the same task, we combine their *representations*. Specifically, we use the encoders of several specialized models as expert feature extractors and fuse their features *before* passing them to a shared decoder.

Mixture of Experts. Increasing the number of parameters generally improves model capacity but also raises inference costs. *Mixture of Experts (MoE)* architectures address this by dividing the model into several smaller sub-networks, or experts, and activating only a subset of them for each input [1]. A routing network learns to select which experts to use, allowing the model to maintain high representational power while reducing computational cost [3]. Our approach shares this modular idea: the encoder is composed of multiple disjoint experts. 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

Our proposed architecture, EoS-FM (Fig. 1), is an ensemble of ConvNeXtV2-Atto [7] encoders (3.4M parameters each). 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. 3.2.

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 Sentinel-1 (S1) and 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 and IR–R–G 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 (as in 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.

These 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.

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 [4]. We hypothesize that such

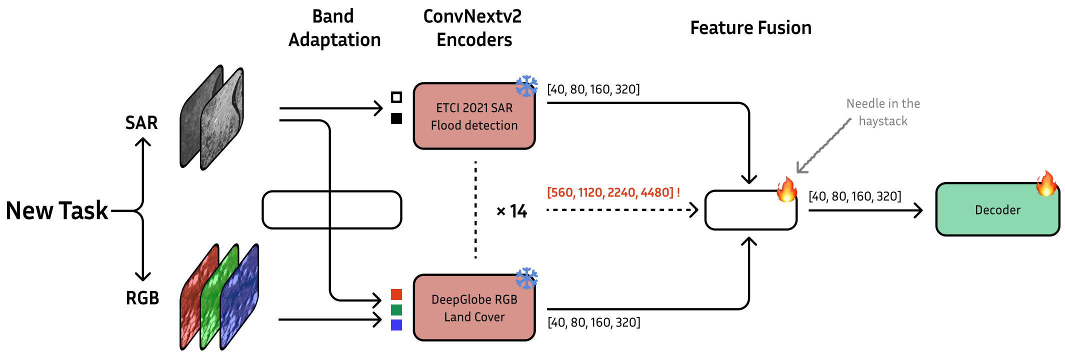

Figure 1: The EoS-FM Backbone adapts any given input to a multitude of formats using band duplication and selection to extract as many feature maps as possible, and then fuse them.

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.

The fused features are passed to a decoder to produce the final output. The ensemble serves as a foundation model: for new downstream tasks, all encoders are frozen, and only the linear fusion and decoder layers are trained. While one could argue that the fusion layer partially overlaps with the encoder, its minimal size means that in practice, equivalent fusion could be implemented within the decoder. We observe similar performance with and without this layer, though the explicit fusion improves efficiency by avoiding extremely wide decoder inputs.

3.2 Training

We train 14 *ConvNeXtV2-Atto* encoders across 9 datasets and 3 modalities (see Appendix A). The datasets were selected to cover a broad range of inputs, including RGB, multispectral, and SAR imagery, as well as different tasks such as classification and segmentation.

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)[1], one using only Sentinel-1 data, and one using only the RGB bands from Sentinel-2 (B4, B3, B2).

Each encoder is initialized from a *ConvNeXtV2-Atto* model pretrained in a self-supervised manner on ImageNet (via the `timm` library [5]), and fine-tuned until convergence on its respective dataset. When input images are too large for deep learning purposes, we use a tiling strategy with a convenient patch size (e.g., 512×512 for MiniFrance).

In total, the Ensemble-of-Specialists Foundation Model (EoS-FM) is trained on 1,373,584 unique samples. Some samples are used multiple times under different modalities due to the modality subsampling strategy, as previously illustrated with BigEarthNet.

4 Experiments

4.1 Downstream Tasks

We follow the setup from the Pangaea Benchmark [11], freezing the EoS-FM encoder ensemble and fine-tuning the UperNet decoder for 80 epochs with a batch size of 8. The best checkpoint is selected based on validation mIoU, and final results are reported using the test mIoU. As shown in Tab. 1, our method achieves performance close to that of much larger models, and even significantly surpasses most RSFMs on the AI4Farms dataset. Using the Distance To Best (DTB) metric [12], we find that our model achieves the best average score, mainly due to its strong performance on AI4Farms.

[1] Encoders with many input bands (e.g., 14) are used only when the data has all the required bands. In the subsequent experiments, only 11 encoders are active due to this.

	HLS Burns	MADOS	Sen1FLoods11	AI4Farms	Mean DTB ↓
RemoteCLIP (87M) ❄	76.59	60.00	74.26	25.12	13.46
GFM-Swin (87M) ❄	76.90	64.71	72.60	27.19	12.10
Prithvi (87M) ❄	83.62	49.98	90.37	26.86	9.74
SatlasNet (87M) ❄	79.96	55.86	90.30	25.13	9.64
DOFA (112M) ❄	80.63	59.58	89.37	27.07	8.30
CROMA (303M) ❄	82.42	**67.55**	90.89	25.65	5.80
EoS-FM (48M, Ours) ❄	79.50	59.90	89.26	45.10	4.17
UNet (~8M) 🔥	**84.51**	54.79	**91.42**	**46.34**	**3.19**

Table 1: Preliminary results on the single temporal datasets of Pangaea benchmark [11]. Although it does not set a new SOTA, our method shows results comparable to larger self-supervised models, only being surpassed by the supervised baseline on average. (DTB = Distance To Best)

4.2 Scaling & Pruning

The results presented in the previous section were obtained with a fixed version of EoS-FM composed of 14 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 study how downstream performance scales with the number of encoders, we performed an ablation experiment where we progressively deactivated encoders by setting their outputs to zero before the feature fusion layer. We then gradually reactivated them until the full ensemble was restored. We conducted this study on the HLS Burn Scars dataset from the Pangaea Benchmark. As shown in Figure 2, the performance of EoS-FM consistently increases with the number of active encoders, showing no clear signs of saturation up to 11 encoders. We plan to further increase the ensemble size to determine where this scaling trend plateaus.

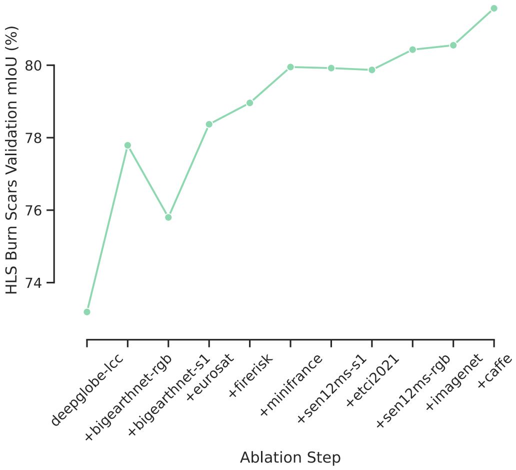

Figure 2: Ablation study: increasing the number of encoders reliably increases the performance of the ensemble in a frozen setting.

The same modular structure offers a simple approach to model pruning: reducing the number of encoders and retaining only those most relevant to a specific downstream task. Preliminary experiments suggest this is an effective strategy. On HLS Burn Scars, we reduced the parameter count of EoS-FM by 57% (from 47.6M to 20.4M) while observing only a –0.38% drop in validation mIoU, by selecting an optimal subset of 6 encoders out of the original 14.

5 Conclusion

In this work, we introduced EoS-FM, an ensemble-based approach to building efficient and modular foundation models for 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 foundation models from smaller, domain-specialized parts rather than relying solely on monolithic architectures. Future work will explore extending the ensemble with additional modalities and tasks, as well as investigating adaptive routing and pruning strategies to further improve efficiency.

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] Robert A Jacobs et al. "Adaptive mixtures of local experts". In: *Neural computation* 3.1 (1991), pp. 79–87.

[2] Thomas G Dietterich. "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15.

[3] Noam Shazeer et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer". In: *arXiv preprint arXiv:1701.06538* (2017).

[4] Amir R Zamir et al. "Taskonomy: Disentangling task transfer learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3712–3722.

[5] Ross Wightman. *PyTorch Image Models*. `https://github.com/rwightman/pytorch-image-models`. 2019. DOI: `10.5281/zenodo.4414861`.

[6] Favyen Bastani et al. "Satlaspretrain: A large-scale dataset for remote sensing image understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 16772–16782.

[7] Sanghyun Woo et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 16133–16142.

[8] Keumgang Cha, Junghoon Seo, and Taekyung Lee. "A Billion-scale Foundation Model for Remote Sensing Images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024), pp. 1–17. ISSN: 1939-1404, 2151-1535. DOI: `10.1109/JSTARS.2024.3401772`. arXiv: `2304.05215 [cs]`. (Visited on 12/16/2024).

[9] Philipe Dias et al. "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*. 2024, pp. 597–600.

[10] Xin Guo et al. "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)*. Seattle, WA, USA: IEEE, June 2024, pp. 27662–27673. ISBN: 979-8-3503-5300-6. DOI: `10.1109/CVPR52733.2024.02613`. (Visited on 01/30/2025).

[11] Valerio Marsocci et al. "PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models". In: arXiv:2412.04204 (Dec. 2024). DOI: `10.48550/arXiv.2412.04204`. arXiv: `2412.04204 [cs]`. (Visited on 12/06/2024).

[12] Pierre Adorni et al. "Towards Efficient Benchmarking of Foundation Models in Remote Sensing: A Capabilities Encoding Approach". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 3096–3106.

[13] Oriane Siméoni et al. "Dinov3". In: *arXiv preprint arXiv:2508.10104* (2025).

[14] Kang Wu et al. "A semantic-enhanced multi-modal remote sensing foundation model for Earth observation". In: *Nature Machine Intelligence* (2025), pp. 1–15.

A Training datasets

Encoder Name	Dataset	Modality	# Bands	# Images	Task	Head
eurosat-s2	EuroSat	S2 MS	13	27,000	Cls.	Linear
caffe	Caffe	SAR	3	681	Seg.	UperNet
sen12ms-s1	Sen12MS	SAR	3	180,662	Seg.	UperNet
sen12ms-s2	Sen12MS	S2 MS	13	180,662	Seg.	UperNet
deepglobe-lcc	DeepGlobe LCC	RGB	3	18,000	Seg.	UperNet
bigearthnet-s1	BigEarthNetV2	SAR	3	549,488	Cls.	Linear
eurosat	EuroSat	RGB	3	27,000	Cls.	Linear
bigearthnet	BigEarthNet	S2 + S1	14	549,488	Cls.	Linear
firerisk	Firerisk	RGB	3	91,872	Cls.	Linear
bigearthnet-rgb	BigEarthNet	RGB	3	549,488	Cls.	Linear
sen12ms-rgb	Sen12MS	RGB	3	180,662	Seg.	UperNet
imagenet	ImageNet	RGB	3	1,281,167	MAE	FCMAE
minifrance	MiniFrance	RGB	3	472,476	Seg.	UperNet
etci2021	ETCI2021	SAR	3	33,405	Seg.	UperNet

Table 2: Summary of the datasets used to train the EoS-FM ensemble. Cls and Seg stand for Classification and Segmentation, respectively.

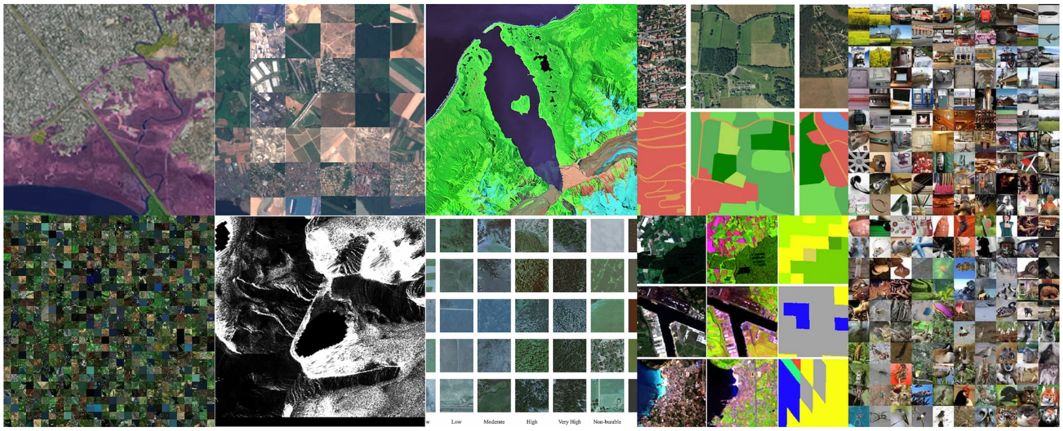

Figure 3: Our EoS-FM model is trained on a very diverse dataset in terms of modalities, resolutions, location, and labels.