Overlap-Free Modality Generalization in Remote Sensing Foundation Models

Gulnaz Zhambulova[1] Yonghao Xu[1] Amanda Berg[1,2] Leif Haglund[1,2] Michael Felsberg[1]

[1]Computer Vision Laboratory, Linköping University, [2]Vantor, Linköping, Sweden

gulnaz.zhambulova@liu.se

Introduction

Understanding how well foundation models generalize across sensing modalities is essential for building truly universal representations in Earth observation. Existing multi-modal or any-sensor models rely on exposure to multiple sensors during pretraining, making evaluation on truly unseen modalities impossible. Therefore, our work isolates pure, overlap-free modality generalization by testing whether optical-only pretraining can transfer to radar without any paired or multi-modal data.

- **Optical imagery** (Sentinel-2): records reflected sunlight across visible to shortwave infrared spectra;
- **Radar data** (Sentinel-1): actively measures microwave backscatter, capturing surface geometry, roughness, and moisture.

These fundamentally different sensing principles make optical–radar transfer a challenging yet informative test of cross-modality generalization.

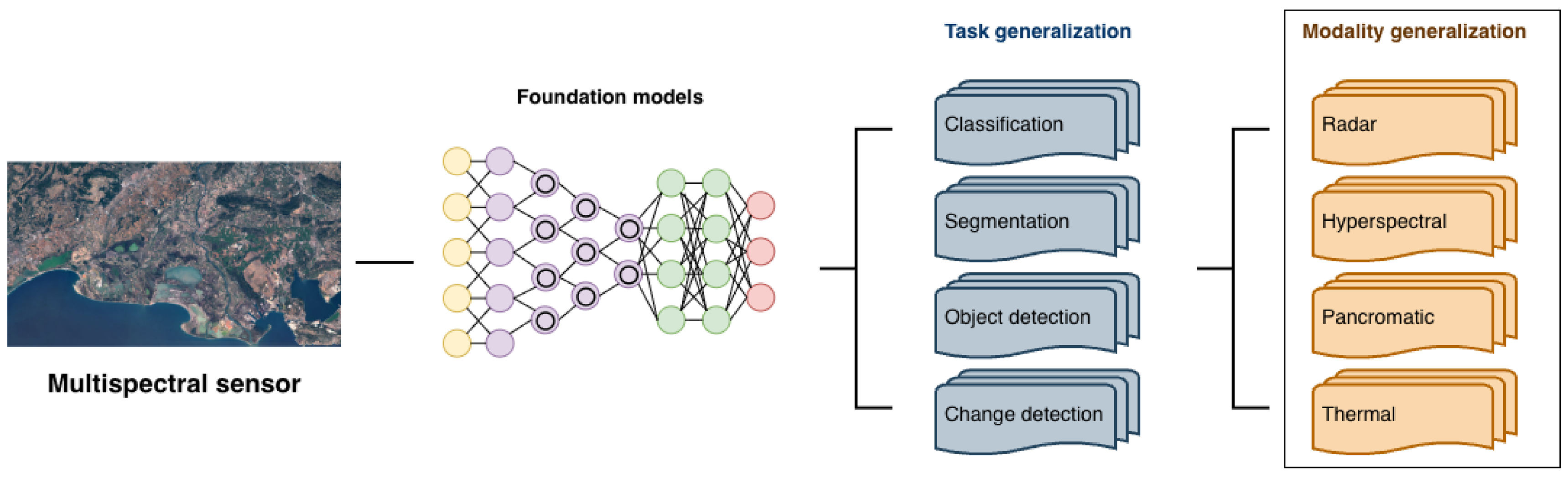

Figure 1: Overview of cross-modality expectations for remote sensing foundation models. A model should ideally produce representations that remain useful for downstream tasks across unseen modalities.

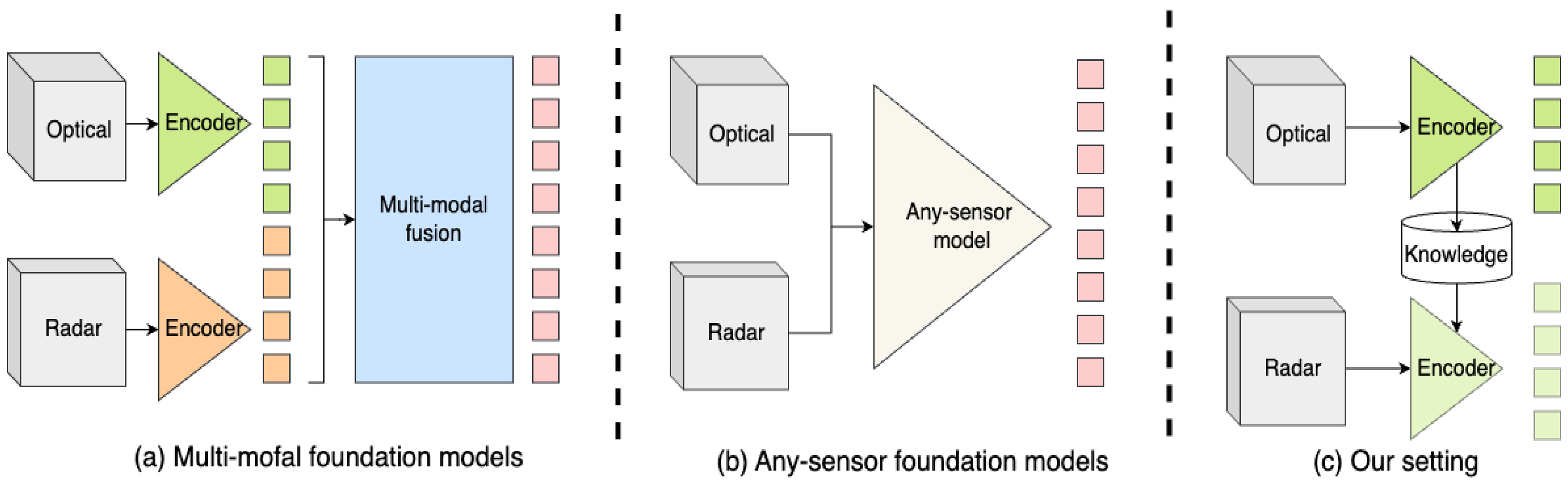

Figure 2: Comparison of modality generalization setups. (a) Multi-modal models use modality-specific encoders with fusion. (b) Any-sensor models share one encoder across modalities. (c) Our overlap-free setting tests a model pretrained on optical directly on radar without paired data.

Methodology

Our design with channel-separated patching, separable positional encodings, and dual reconstruction losses aims to capture spatial-spectral structure robust enough to generalize across sensing principles. Our foundation model adopts the Masked Autoencoder (MAE) framework [1], where the network learns to reconstruct missing image patches from a partially observed input. Motivated by ChannelViT [2], we adopt a patch size of (1, 16, 16), treating each spectral band as a separate channel token.

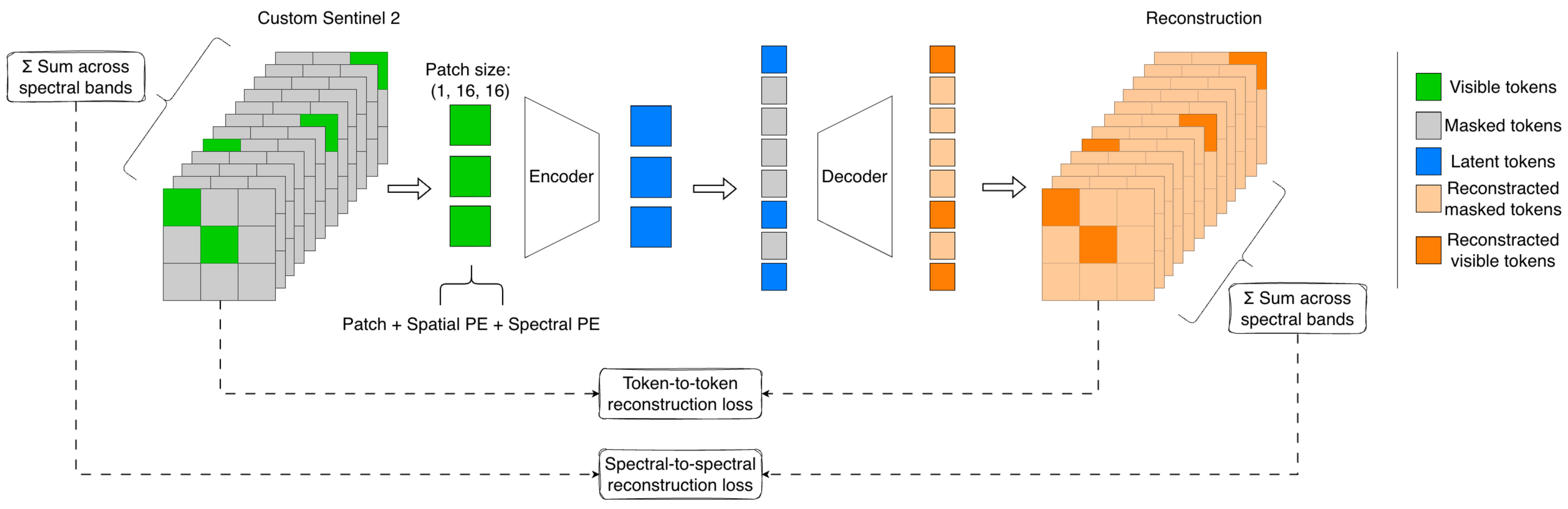

Figure 3: MAE pipeline with token-to-token and spectral-to-spectral reconstruction losses for optical pretraining.

To disentangle spatial and spectral features, we use separable positional embeddings as in [3]. Specifically, the spatial positional embedding encodes the location of each patch within the 2D spatial grid, while the spectral positional embedding encodes which spectral segment each patch belongs to. Following [4], the total loss is defined as the sum of token-to-token and spectral-to-spectral reconstruction losses:

$$\mathcal{L}_{\text{spectral-to-spectral}} = \frac{1}{H_p \cdot W_p} \sum_{h=1}^{H_p} \sum_{w=1}^{W_p} \|\hat{Y}^{\text{spatial}}_{h,w} - Y^{\text{spatial}}_{h,w}\|^2$$

$$\mathcal{L}_{\text{token-to-token}} = \frac{1}{\sum_i M_i} \sum_i M_i \, \|\hat{Y}_i - Y_i\|^2$$

where $M_i \in \{0, 1\}$ indicates whether patch i is masked or visible, Y_i denotes the ground-truth patch, and Y'_i its reconstruction.

Dataset: For pretraining, a custom Sentinel-2 dataset is used, constructed to ensure globally balanced coverage across diverse land-cover types. For downstream task, the BigEarthNet-MM (BEN) dataset [5] is used.

Experimental Results

To estimate evaluation variance within computational limits, the test set was randomly divided into five groups and repeated three times with different seeds; mean and standard deviation were then computed across all subsets.

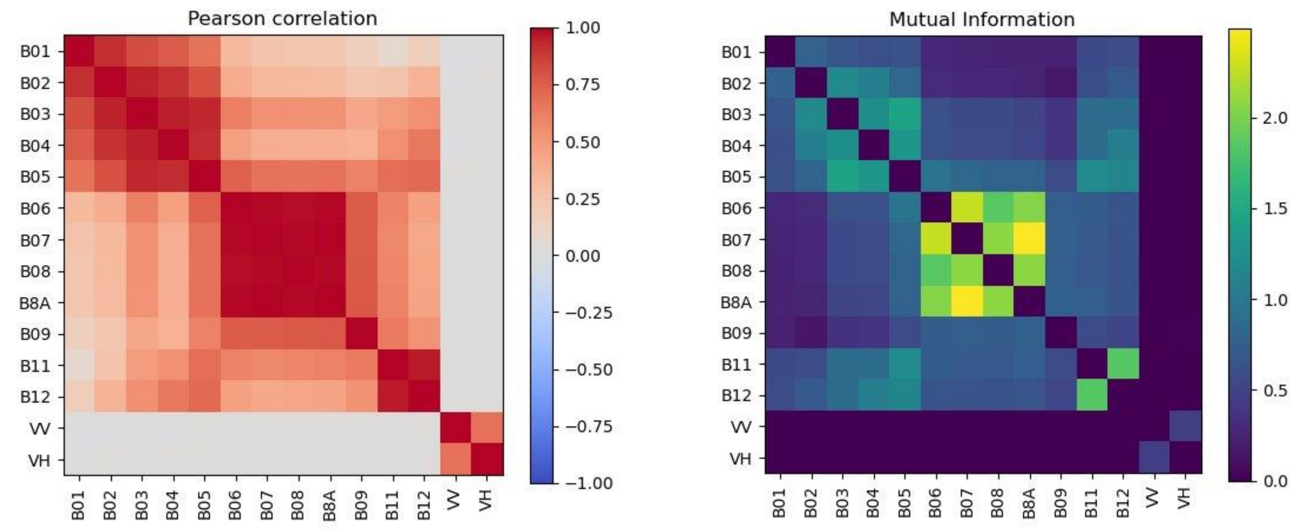

Figure 4: Correlation and mutual information between Sentinel-2 spectral bands and Sentinel-1 radar channels. Low correlation and mutual information between optical and radar channels illustrate the challenge of cross-modality transfer.

Table 1: Comparison of pretraining sources and patch sizes for cross-modality (Sentinel-2 -> Sentinel-1) and within-modality (Sentinel-2 -> Sentinel-2) transfer.

Pretraining	Patch size	BEN Sentinel-1		BEN Sentinel-2	
		mAP ↑	F1 ↑	mAP ↑	F1 ↑
From scratch	(1, 16, 16)	68.47 ± 0.13	56.82 ± 0.14	78.09 ± 0.05	67.79 ± 0.07
Sentinel-2	(1, 16, 16)	$\mathbf{71.02 \pm 0.13}$	59.18 ± 0.15	81.18 ± 0.08	70.78 ± 0.11
RGB	(1, 16, 16)	70.83 ± 0.14	58.93 ± 0.14	80.25 ± 0.07	69.91 ± 0.10
Sentinel-2	(1, 8, 8)	70.81 ± 0.14	$\mathbf{59.53 \pm 0.17}$	$\mathbf{82.35 \pm 0.09}$	$\mathbf{72.14 \pm 0.09}$

Optical pretraining consistently outperforms training from scratch on Sentinel-1 by more than two points in mAP and F1, demonstrating that modality-agnostic structural priors can emerge from single-modality training.

Table 2: Comparison of different masking and reconstruction strategies. S1: random 3D patch masking; S2: input and reconstruct one random band; S3: input one band, reconstruct remaining bands; S4: input three random bands, reconstruct remaining bands.

ID	BEN Sentinel-1		BEN Sentinel-2	
	mAP ↑	F1 ↑	mAP ↑	F1 ↑
S1	$\mathbf{71.02 \pm 0.13}$	$\mathbf{59.18 \pm 0.15}$	$\mathbf{81.18 \pm 0.08}$	$\mathbf{70.78 \pm 0.11}$
S2	69.92 ± 0.13	58.15 ± 0.17	78.66 ± 0.08	68.01 ± 0.10
S3	68.42 ± 0.11	56.46 ± 0.15	74.73 ± 0.10	63.61 ± 0.12
S4	70.25 ± 0.13	58.76 ± 0.12	71.19 ± 0.11	61.24 ± 0.11

Table 3: Comparison of partial fine-tuning strategies for cross-modality transfer. Only selected components of the pretrained model were updated during fine-tuning.

Trainable components	BEN Sentinel-1	
	mAP ↑	F1 ↑
Head only	56.58 ± 0.14	55.08 ± 0.14
Patch embedding + head	57.22 ± 0.15	55.67 ± 0.15
Patch embedding (reinitialized) + head	56.81 ± 0.16	54.95 ± 0.14
Patch embedding + spectral encoding + head	$\mathbf{57.51 \pm 0.15}$	$\mathbf{55.87 \pm 0.15}$

Random 3D patch masking provides the strongest transferable features, and only full fine-tuning unlocks substantial radar performance, highlighting both the promise and limits of optical-to-radar generalization.

Conclusion

We presented the first systematic study of strict cross-modality transfer between optical and radar domains using a single-modality pretrained foundation model. Results show that masked autoencoder pretraining on Sentinel-2 improves Sentinel-1 performance without any radar exposure, indicating that structural and physical priors learned from optical data extend beyond their original modality. While partial fine-tuning offers limited adaptation, full fine-tuning remains necessary for strong radar transfer. Future work will explore reverse transfer (radar->optical) and compare with multi-modal or any-sensor foundation models to further understand the limits of modality-agnostic representation learning.

References

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

[2] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: an image is worth 1 x 16 x 16 words. arXiv preprint arXiv:2309.16108, 2023.

[3] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24088–24097, 2024.

[4]Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Pedram Ghamisi, Naoto Yokoya, Hao Li, 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, (8):5227–5244, 2024.

[5] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. IEEE Geoscience and Remote Sensing Magazine, 9(3):174–180, 2021