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]

Abstract

Understanding how well foundation models generalize across sensing modalities is essential for building truly universal representations in Earth observation. However, it remains unclear whether single-modality pretraining can generalize across fundamentally different sensing principles. This work presents the first systematic study of overlap-free cross-modality transfer from optical (Sentinel-2) to radar (Sentinel-1) imagery. A masked autoencoder pretrained solely on Sentinel-2 is evaluated on Sentinel-1 without radar exposure. Optical pretraining consistently surpasses training from scratch by over two points in mAP and F1, providing empirical evidence that optical pretraining enables overlap-free modality generalization.

1 Introduction

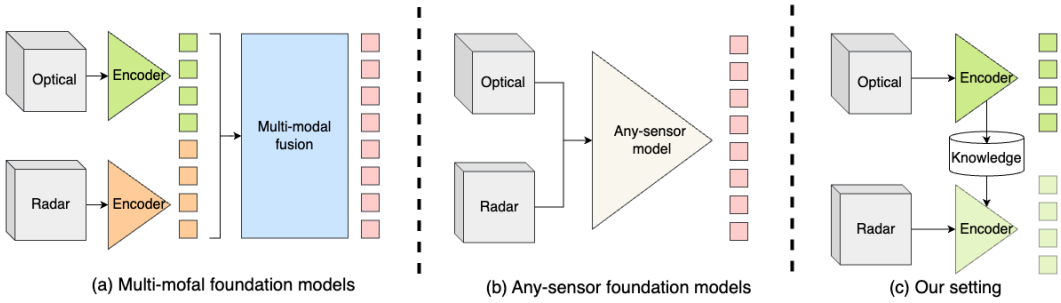

Figure 1: Comparison of modality generalization setups. (a) Multi-modal models use modality-specific encoders with fusion. (b) Any-sensor models share one encoder across modalities. (c) Our overlap-free setting tests a model pretrained on optical directly on radar without paired data.

Remote sensing imagery differs substantially from natural images due to its non–object-centric nature, sensor variability, and diverse spatial, spectral, and temporal resolutions [1], motivating the need for specialized models. While satellite data volumes continue to grow, generating reliable annotations remains labor-intensive and requires expert knowledge. To address this, Remote Sensing Foundation Models (RSFMs) leverage large collections of unlabelled imagery through self-supervised pretraining, achieving strong performance across diverse downstream tasks and consistently outperforming ImageNet-pretrained counterparts [2, 3, 4, 5, 6].

RSFMs are expected to generalize across tasks, geographies, and sensors. While the first two aspects have been explored through benchmarks covering diverse regions and applications, sensor generalization remains an open challenge. Recent studies have begun to address this issue by developing

[*]Corresponding author, [1]Computer Vision Laboratory, Linköping University, [2]Vantor, Linköping, Sweden. Emails: {firstname.lastname}@liu.se

multi-modal RSFMs that jointly learn from data captured by different sensors. These models typically fuse features extracted from sensor-specific encoders or tokenizers using geographically aligned images, improving performance on both single- and multi-modal downstream tasks [7, 8, 9, 10, 11, 12]. However, as these methods focus on exploiting the complementary nature of known modalities rather than achieving generalization to unseen ones, downstream tasks are still limited to the modalities used during pretraining [13].

To overcome this challenge, recent works on any-sensor RSFMs have shifted toward sensor-independent architectures that can adapt to varying numbers of input channels, thereby enabling compatibility with different modalities. This flexibility is achieved through wavelength-aware dynamic patch embedding [13, 14], channel-wise tokenization enriched with spectral encoding [15, 16], or modality- and spectrum-aware projection layers [17, 18]. However, these models are typically pretrained on multi-modal datasets that already cover the most common remote sensing modalities, such as high-resolution RGB, multispectral, hyperspectral, and SAR, making evaluation on truly unseen modalities infeasible. Attempts to simulate unseen configurations often rely on derived or optical-only datasets, which remain within overlapping spectral ranges and thus reflect intra-domain rather than true cross-modality generalization.

Motivated by this limitation, this work investigates an open question in the modality generalization of RSFMs: *can a foundation model pretrained on one modality (optical imagery) provide transferable representations for a completely unseen one (radar)?* Optical imagery (Sentinel-2) records reflected sunlight across visible to shortwave infrared spectra [19], while radar data (Sentinel-1) actively measures microwave backscatter, capturing surface geometry, roughness, and moisture [20]. These fundamentally different sensing principles make optical–radar transfer a challenging yet informative test of cross-modality generalization. Rather than learning explicit cross-modal correspondences, our approach directly evaluates the intrinsic generalization ability of RSFMs. Understanding such cross-modality transfer is essential for developing universal RSFMs that capture physical invariances and remain reliable in data-scarce or single-sensor scenarios. To the best of our knowledge, this is the first systematic study to investigate strict cross-modality transfer between optical and radar domains using a single-modality foundation model.

2 Methodology

Masked autoncoder. Our foundation model adopts the Masked Autoencoder (MAE) framework [21], where the network learns to reconstruct missing image patches from a partially observed input. Unlike the standard MAE setup used for RGB imagery, where patches span both spatial and channel dimensions, we adopt a patch size of $(1,\ 16,\ 16)$, treating each spectral band as a separate channel token. This design choice is motivated by ChannelViT [22], which showed that decoupling the spectral and spatial dimensions enables transformers to learn richer inter-channel relationships. Moreover, this design allows flexibility in the number of input channels the model can accommodate.

Positional embedding. To disentangle spatial and spectral features, we use separable positional embeddings as in [23]. Specifically, the spatial positional embedding encodes the location of each patch within the 2D spatial grid as $\texttt{pos\_embed\_spatial} \in \mathbb{R}^{1\times H\cdot W\times D}$, where $H \cdot W$ denotes the number of spatial patches, and D is the embedding dimension. Similarly, the spectral positional embedding encodes which spectral segment each patch belongs to as $\texttt{pos\_embed\_spectral} \in \mathbb{R}^{1\times C\times D}$, where C is the number of spectral segments. The final positional encoding is the sum of both embeddings.

Loss function. Following [6], the total loss is defined as the sum of token-to-token and spectral-to-spectral reconstruction losses. For the token-to-token reconstruction loss, we reconstruct only the masked patches:

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

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

To further enforce spectral consistency, a spectral-to-spectral reconstruction loss is introduced. Specifically, both Y and $\hat{Y}$ are first aggregated across the spectral dimension as $Y^{\text{spatial}}_{h,w} = \sum_{c=1}^{C} Y_{c,h,w}$,

and $\hat{Y}^{\text{spatial}}_{h,w} = \sum_{c=1}^{C} \hat{Y}_{c,h,w}$, where C is the number of spectral channels, and (h, w) indexes spatial patch locations. The spectral-to-spectral reconstruction loss is then defined as:

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

where H_p and W_p denote the number of patch positions along height and width.

3 Experimental settings

Pre-training dataset. A custom Sentinel-2 dataset was constructed to ensure globally balanced coverage across diverse land-cover types. Regions of Interest (ROIs) were sampled using the ESA WorldCover 2020 map [24], which defines 11 standardized land-cover classes. Unlike prior urban- or biome-focused sampling strategies [2, 5], our approach targets uniform representation of built-up, agricultural, and natural environments. Sampling was performed per continent and subdivided into a 10×10 spatial grid in Google Earth Engine [25], selecting a fixed number of points per class with a minimum spacing of 3 km to avoid overlap. Cloud-free Sentinel-2 composites were generated by mosaicking all images from a randomly chosen month (primarily 2023, with 2021–2022 fallback if cloud cover exceeded 20%). Sample counts per grid cell were scaled by land area to maintain global class balance. For computational efficiency, experiments were conducted on a 10% stratified subset (93K images) balanced across continents and land-cover classes.

Downstream task. The BigEarthNet-MM (BEN) dataset [26] serves as the primary downstream benchmark, providing Sentinel-1 and Sentinel-2 image pairs across ten European countries. It helps with evaluation of cross-modality transfer, with Sentinel-2 used for within-modality and Sentinel-1 for unseen-modality testing. Each of the 590,326 images is annotated with 19 multi-label land-cover classes. Following prior work [2, 6], 10% of the training set is used for fine-tuning.

Pretraining setup. We apply random spatial masking independently to each spectral band, following the MAE protocol, with 90% of patch tokens masked per iteration. Images are normalized per band using reflectance scaling and augmented with random flips and crops. Models are pretrained for 100 epochs using the AdamW optimizer (learning rate 1×10^{-4}, 10-epoch warmup, weight decay 0.05) with a ViT-Base encoder and a 4-block, 512-dimensional decoder [27, 21]. Training is performed on eight NVIDIA A100 GPUs with mixed precision.

Cross-modality transfer evaluation. Cross-modality transfer was evaluated using two adaptation strategies: full fine-tuning and partial fine-tuning. Full fine-tuning updated all model parameters, whereas partial fine-tuning froze the encoder and trained only selected modality-related components to assess representation generalization. Both were trained for 30 epochs under identical settings. Performance is reported as mean Average Precision (mAP) and F1-score. 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.

4 Results and discussion

Cross-modality evaluation. As shown in Table 1, self-supervised pretraining on Sentinel-2 imagery consistently improves performance for both within-modality and cross-modality transfer. The Sentinel-2–pretrained model surpasses the from-scratch baseline by more than two points in both mAP and F1 on Sentinel-1, demonstrating that optical pretraining learns features transferable to radar data despite no radar exposure. Differences between Sentinel-2 and RGB pretraining are minor and fall within the measured variance, indicating that spectral diversity provides only a modest benefit. The gains over the from-scratch baseline, however, remain significant. For within-modality transfer, the smaller patch size $(1, 8, 8)$ achieves the highest scores, indicating that finer spatial granularity can help model optical structures, though gains are modest relative to the computational overhead.

Effect of masking and reconstruction strategies. Table 2 shows that the standard 3D patch masking (`S1`) yields the highest accuracy across both modalities, confirming that joint spatial–spectral

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	$\mathbf{59.53 \pm 0.17}$	$\mathbf{82.35 \pm 0.09}$	$\mathbf{72.14 \pm 0.09}$

masking promotes more transferable representations. Band-wise variants (S2–S4) perform worse, especially on Sentinel-2, suggesting that reduced spectral context limits inter-band learning. Nevertheless, all pretrained variants surpass the from-scratch baseline on Sentinel-1, indicating that even partial spectral masking supports spatial transfer.

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

Effect of partial fine-tuning. Table 3 shows that training only the classification head provides limited adaptation, while including the patch embedding slightly improves cross-modality transfer. Reinitializing the embedding reduces performance, confirming the value of pretrained spatial filters. The best partial setup involves updating the patch embedding, spectral encoding, and head, yielding modest but consistent gains. However, all partial strategies remain notably below full fine-tuning performance, indicating that broader parameter adaptation is required for effective optical-to-radar generalization.

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	$\mathbf{57.51 \pm 0.15}$	$\mathbf{55.87 \pm 0.15}$

5 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] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

[2] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.

[3] 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.

[4] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[5] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024.

[6] 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.

[7] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[8] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[9] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 27852–27862, 2024.

[10] Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. Skysense v2: A unified foundation model for multi-modal remote sensing. *arXiv preprint arXiv:2507.13812*, 2025.

[11] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. Terrafm: A scalable foundation model for unified multisensor earth observation. *arXiv preprint arXiv:2506.06281*, 2025.

[12] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

[13] 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 preprint arXiv:2403.15356*, 2024.

[14] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. *arXiv preprint arXiv:2503.11849*, 2025.

[15] Jonathan Prexl and Michael Schmitt. Senpa-mae: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *DAGM German Conference on Pattern Recognition*, pages 317–331. Springer, 2024.

[16] Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2204–2214, 2025.

[17] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19530–19540, 2025.

[18] Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. Smarties: Spectrum-aware multi-sensor auto-encoder for remote sensing images. *arXiv preprint arXiv:2506.19585*, 2025.

[19] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[20] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012.

[21] 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.

[22] 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.

[23] 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.

[24] Daniele Zanaga, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, Alex Grosu, Audrey Paccini, Sylvain Vergnaud, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Ivelina Georgieva, Myroslava Lesiv, Sarah Carter, Martin Herold, Linlin Li, Nandin-Erdene Tsendbazar, Fabrizio Ramoino, and Olivier Arino. Esa worldcover 10 m 2020 v100, October 2021.

[25] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

[26] 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.

[27] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.