GEOCROSSBENCH: CROSS-BAND GENERALIZATION FOR REMOTE SENSING

Hakob Tamazyan, Ani Vanyan, Alvard Barseghyan, Anna Khosrovyan & Hrant Khachatrian
YerevaNN
Yerevan State University
Yerevan, Armenia
{hakob,ani,alla,anna,hrant}@yerevann.com

Evan Shelhamer
University of British Columbia
Vector Institute
Vancouver, BC & Toronto, ON, Canada
shelhamer@cs.ubc.ca

ABSTRACT

The number and diversity of remote sensing satellites grows over time, yet the vast majority of labeled data comes from older satellites. As foundation models for Earth observation scale up, the cost of (re-)training to support new satellites grows too, making spectral generalization critical. We introduce GeoCrossBench, an extension of GeoBench with a new evaluation protocol: it tests in-distribution performance; generalization to satellites with no band overlap; and generalization to satellites with additional bands. We also develop a self-supervised extension of ChannelViT, χViT, to improve cross-satellite performance. We show that while even the best RS foundation models do not outperform general-purpose models like DINOv3 in our benchmark, our χViT outperforms the runner-up DINOv3. Finally, we show that performance of all tested models drops by 5-25% when given additional bands during test time, highlighting that current architectures are not yet future-proof.

1 INTRODUCTION

The growth of remote sensing (RS) data has led to sophisticated deep learning models capable of analyzing complex geospatial patterns. Pre-trained foundation models have emerged as a popular paradigm for learning generalizable representations (Xiong et al., 2024; Fuller et al., 2023; Jakubik et al., 2025; Cong et al., 2022). However, RS data is inherently multimodal, capturing diverse spectral *bands* including multispectral, hyperspectral, and synthetic aperture radar (SAR).

While recent foundation models transfer well when train and test bands match, their **cross-band generalization**—to bands and sensors unseen during fine-tuning—remains limited. This is a critical gap: a practitioner might need to transfer a model trained on public Sentinel-2 data to newer platforms like Planet SuperDove (which introduces new bands) or SatVu's HotSat (thermal data). The most extreme case involves transferring between modalities, such as optical to SAR.

We introduce **GeoCrossBench** to assess this gap with three protocols: (1) *in-distribution*, (2) *no overlap bands*, and (3) *superset bands*. We evaluate a range of existing foundation models and introduce a new baseline, χ**ViT** (ChiViT), a ChannelViT (Bao et al., 2024) variant pretrained with iBOT (Zhou et al., 2022) on a large-scale multi-modal dataset. Our findings underscore the pressing need for rigorous benchmarks to drive the development of versatile Earth observation models.

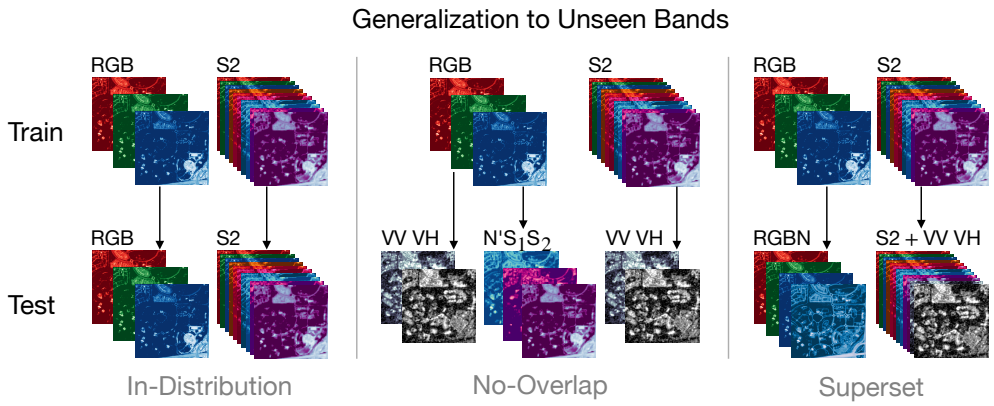


Figure 1: The GeoCrossBench evaluation framework. (1) *In-Distribution*: Fine-tune and evaluate on identical bands. (2) *No-Overlap*: Evaluate transfer to disjoint bands (e.g., RGB→SAR). (3) *Superset*: Evaluate on strict supersets of training bands (e.g., RGB→RGB+NIR).

2 Related Work

Foundation Models for Remote Sensing. Recent work has established the utility of foundation models for RS. SatMAE (Cong et al., 2022) demonstrated self-supervised pre-training at ViT scale, while Satlas (Bastani et al., 2023) explored large-scale supervised pre-training. Scale-MAE (Reed et al., 2023) addressed generalization across spatial resolutions. GeoCrossBench complements these by focusing on *spectral* generalization.

Multi-modal Learning. Existing works explore learning *from* multiple bands but often do not address extending to *new* bands. Models like SoftCon (Wang et al., 2024) and DOFA (Xiong et al., 2024) learn intra-modal representations, while CROMA (Fuller et al., 2023) and TerraFM (Danish et al., 2025) jointly learn inter-modal representations. However, these usually require specific input configurations. Our work evaluates the capability to generalize to unseen spectral configurations without expensive retraining.

3 GeoCrossBench: Dataset and Protocol

GeoCrossBench extends the GeoBench (Lacoste et al., 2023) framework by fusing existing datasets with Sentinel-1 (SAR) data and introducing new tasks to ensure broad spectral coverage.

3.1 Datasets

We utilize Sentinel-2 (10 optical bands) and Sentinel-1 (2 SAR bands: VV, VH) across all datasets. The benchmark covers three core tasks: scene classification, semantic segmentation, and change detection. Table 1 details the datasets, including new additions like x-sen1floods11 and x-harvey tailored for cross-band evaluation. For datasets missing SAR in their original release (e.g., OSCD, EuroSAT), we fused spatially and temporally aligned Sentinel-1 imagery.

3.2 Evaluation Protocols

We define three settings to probe different aspects of generalization (see Figure 1):

1. In-Distribution: Models are trained and evaluated on identical bands.

- *RGB → RGB:* Standard visual fine-tuning baseline.
- *S2 → S2:* Full 10-band multispectral fine-tuning.

2. No-Overlap Bands: A challenging setting testing transfer to completely distinct sensors.

- *RGB → S1* and *S2 → S1:* Generalization from Optical to SAR.

Table 1: Overview of the datasets included in GeoCrossBench. The ones marked with ⋆ are not part of the original GeoBench.

Dataset Name	Image Size	#Classes	Sensors/Bands	Train	Val	Test
			Classification			
x-bigearthnet	120×120	43	S2 (10) + S1 (2)	20000	1000	1000
x-so2sat	32×32	17	S2 (10) + S1 (2)	19992	986	986
x-brick-kiln	64×64	2	S2 (10) + S1 (2)	15063	999	999
x-eurosat	64×64	10	S2 (10) + S1 (2)	2000	1000	1000
			Semantic Segmentation			
x-cashew-plantation	256×256	7	S2 (10) + S1 (2)	1350	400	50
x-SA-crop-type	256×256	10	S2 (10) + S1 (2)	3000	1000	1000
x-harvey-building ⋆	256×256	2	S2 (10) + S1 (2)	375	94	461
x-sen1floods11 ⋆	512×512	2	S2 (10) + S1 (2)	252	89	90
			Change Detection			
x-harvey-flood ⋆	256×256	2	S2 (10) + S1 (2)	375	94	461
x-oscd ⋆	224×224	2	S2 (10) + S1 (2)	24 cities	14 cities	10 cities

- *RGB* $\rightarrow$ *N'S_1S_2:* Transfer to Near-Infrared and SWIR bands (S2 B8A, B11, B12).

3. Superset Bands: Tests robustness when provided with *more* bands at test time than during training.

- *RGB* $\rightarrow$ *RGBN:* Adding Near-Infrared (B8).
- *S2* $\rightarrow$ *S2+S1:* Fused Optical+SAR inference after Optical-only training.

4 χViT: A New Baseline for Cross-Band Transfer

We extend ChannelViT (Bao et al., 2024) into χ**ViT** (ChiViT), using a hierarchical pretraining recipe.

Architecture. Standard ViTs tokenize a patch of size $P \times P \times C$ into a single token. In contrast, χViT tokenizes each single-channel patch $P \times P \times 1$ independently. We apply a learnable *channel embedding* $\mathbf{e}_c^{\text{chn}}$ in addition to the positional embedding. The input sequence to the Transformer encoder becomes $[\mathbf{e}^{\text{CLS}}; \ldots; Wx_{c,j} + \mathbf{e}_j^{\text{pos}} + \mathbf{e}_c^{\text{chn}}; \ldots]$. The projection W is shared across channels.

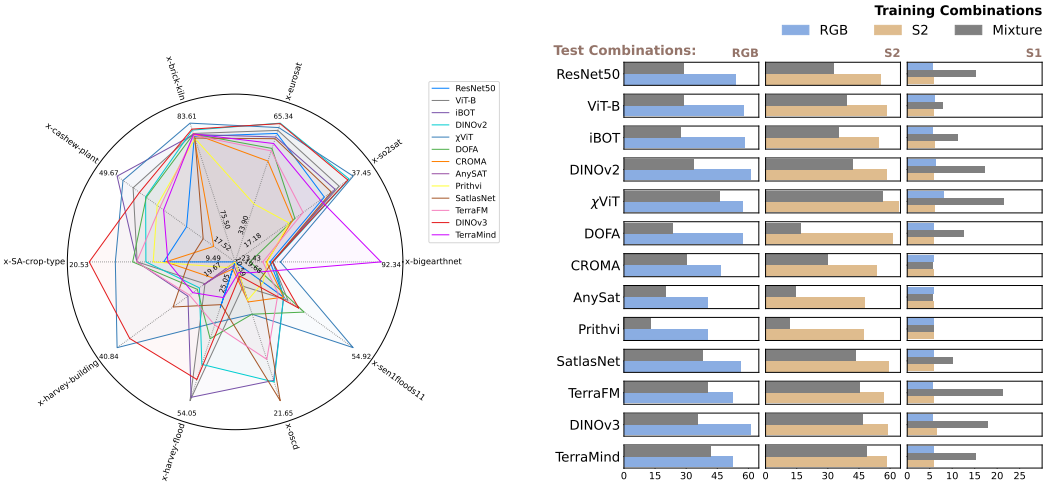
Pretraining. We pretrain using the iBOT (Zhou et al., 2022) paradigm on over 23 million images. We employ *Hierarchical Channel Sampling* during pretraining, where the model sees varying subsets of bands, forcing it to learn robust representations that do not rely on the presence of the full spectrum.

5 Experiments and Discussion

We evaluate specialized RS foundation models (TerraMind, DOFA, SatlasNet, etc.) against general-purpose vision models (DINOv3, DINOv2). Table 2 and Figure 2 summarize the results.

In-Distribution: General-Purpose Models prevail. Even top RS models fail to consistently outperform general-purpose vision models like DINOv3 in the *In-Distribution* setting. This suggests that the scale of pretraining (billions of natural images) provides a feature foundation that current domain-specific models, despite their specialized pretraining, have not yet surpassed for standard RGB or optical tasks.

No-Overlap: χViT shines. The limitations of current models appear when generalizing to unseen bands. In the *No-Overlap* setting (e.g., S2 $\rightarrow$ S1), all models suffer a severe 2-4x drop in performance. However, χViT significantly outperforms all competitors here (see Figure 2a). By tokenizing channels independently, χViT learns to extract value from SAR or SWIR bands based on their



(a) Radar plot for all models and datasets we tried.

(b) Linear probing results on x-so2sat.

Figure 2: **Performance summary on GeoCrossBench.** (a) Radar chart of model performance. (b) Linear probing on x-so2sat shows that accessing oracle bands (mixture) improves performance on the S1 split.

specific embeddings, rather than failing because the input distribution (channel count/stats) has shifted fundamentally.

Superset: Adding bands hurts. Counter-intuitively, providing models with *more* information at test time (e.g., RGB $\rightarrow$ RGBN) degrades performance by 5-25%. Current architectures appear to overfit to the training band configuration; they treat the additional band as a noise source rather than a signal, lacking the mechanism to integrate it zero-shot.

Is the benchmark saturated? To verify that the poor performance on S1 isn't simply due to a lack of signal in radar data, we performed a linear probing experiment (Fig. 2b). We trained a linear probe

Table 2: Performance evaluation of all tested models on GeoCrossBench. The * symbol indicates the frozen backbone. The performance metrics of each setting represent the average scores across all GeoCrossBench datasets.

	In-Distribution				No-Overlap					Superset				
Fine-tuned on **Tested on**	**RGB** **RGB**	**S2** **S2**	**AVG**	**#**	**RGB** **S1**	**S2** **S1**	**RGB** **N'S_1S_2**	**AVG**	**#**	**RGB** **RGBN**	**S2** **S2+S1**	**AVG**	**#**	**Overall** **AVG**
χViT	61.81	63.53	62.67	6	17.96	20.93	30.37	23.09	3	59.03	57.96	58.49	1	44.51
DINOv3	62.46	63.0	62.73	5	17.62	17.19	27.76	20.86	5	52.48	59.62	56.05	2	42.88
iBOT	64.73	61.73	63.23	2	18.83	14.63	26.95	20.13	8	52.36	57.04	54.7	3	42.32
TerraMind	57.78	66.32	62.05	8	28.1	24.94	28.82	27.29	1	49.97	34.4	42.18	15	41.48
ViT-B	62.77	62.72	62.75	4	18.87	14.75	25.18	19.6	10	46.03	58.23	52.13	4	41.22
DINOv2	65.26	62.53	63.89	1	17.36	15.01	25.85	19.41	11	54.52	47.59	51.06	5	41.16
χViT*	56.95	58.42	57.69	13	19.02	18.92	27.12	21.69	4	47.75	48.6	48.17	8	39.54
DINOv2*	61.77	56.58	59.17	10	16.28	15.84	30.13	20.75	7	51.13	38.86	45.0	10	38.66
DOFA	61.71	64.39	63.05	3	17.34	11.77	13.12	14.08	23	50.53	48.62	49.57	6	38.21
TerraFM	61.85	62.35	62.1	7	15.9	13.53	20.82	16.75	14	40.76	50.23	45.5	9	37.92
TerraMind*	52.36	61.63	57.0	14	23.26	22.38	28.74	24.79	2	44.63	30.32	37.47	20	37.62
SatlasNet	49.23	69.12	59.18	9	14.62	14.61	15.4	14.88	18	38.88	58.58	48.73	7	37.21
iBOT*	62.02	50.9	56.46	16	15.3	13.94	30.08	19.78	9	48.63	38.13	43.38	12	37.0
ResNet50	60.43	57.35	58.89	11	13.29	13.48	19.69	15.48	16	41.53	48.33	44.93	11	36.3
DINOv3*	58.51	49.0	53.75	19	14.91	17.17	30.37	20.81	6	46.69	33.54	40.11	18	35.74
DOFA*	59.1	58.03	58.57	12	15.23	14.46	14.83	14.84	19	38.77	45.05	41.91	16	35.07
TerraFM*	56.7	56.72	56.71	15	14.91	12.16	24.81	17.29	12	40.94	35.24	38.09	19	34.5
CROMA	51.58	56.5	54.04	18	16.18	12.26	16.71	15.05	17	34.04	51.61	42.82	14	34.13
ViT-B*	53.42	50.02	51.72	21	16.18	14.03	21.65	17.29	13	42.87	39.84	41.35	17	34.0
Prithvi	52.61	56.88	54.74	17	13.96	11.87	14.71	13.52	24	32.82	53.04	42.93	13	33.7
ResNet50*	53.13	50.22	51.68	22	12.33	13.44	17.81	14.53	21	41.61	31.04	36.32	21	31.37
SatlasNet*	43.6	51.74	47.67	23	12.18	13.5	13.45	13.04	25	24.9	37.17	31.04	22	28.08
CROMA*	40.99	49.7	45.34	24	14.57	13.0	15.6	14.39	22	20.14	37.48	28.81	23	27.35
AnySAT	47.34	58.81	53.08	20	16.19	13.79	18.3	16.09	15	14.72	13.77	14.24	25	26.13
Prithvi*	43.96	28.39	36.17	26	12.68	10.04	13.29	12.0	26	30.69	26.05	28.37	24	23.58
AnySAT*	40.35	47.01	43.68	25	13.35	13.57	17.52	14.81	20	13.0	12.64	12.82	26	22.49

on a mixture of frozen representations from all available bands (RGB, S2, S1, etc.) on x-so2sat. The result was a significant performance boost on the S1 test set compared to models trained only on S2. This confirms that the useful information exists in the data and the frozen representations, but standard fine-tuning paradigms fail to access it when the modality changes.

6 Conclusion

GeoCrossBench reveals that while remote sensing foundation models are maturing, they remain brittle to spectral shifts. General-purpose models are surprisingly effective in-distribution, but specialized architectures like χViT are required to handle the distinct challenge of cross-band generalization. We release this benchmark to encourage the development of future-proof models that can adapt to the ever-expanding constellation of Earth observation sensors.

References

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. 2024.

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*, pp. 16772–16782, 2023.

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. volume 35, pp. 197–211, 2022.

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.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. volume 36, pp. 5506–5538, 2023.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023.

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*, pp. 4088–4099, 2023.

Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Multi-label guided soft contrastive learning for efficient earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

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.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.