GeoCrossBench: Cross-Band Generalization for Remote Sensing

Hakob Tamazyan [1 2] **Ani Vanyan** [1 2] **Alvard Barseghyan** [1 2] **Anna Khosrovyan** [1 2] **Evan Shelhamer** [3 4] **Hrant Khachatrian** [1 2]

Abstract

Foundation models are transforming Earth observation, yet struggle to generalize across bands and sensors to handle the data for different applications. We introduce GeoCrossBench, a novel benchmark that extends the standard GeoBench, to evaluate this critical *cross-band* capability in remote sensing foundation models. We measure generalization by augmenting datasets with additional optical and radar data, training on RGB, then testing on other bands. We first evaluate existing models, as a reality check on current performance and for analysis of pretraining effects, then evaluate our own self-supervised extension of the ChannelViT model, χViT, to improve cross-band performance. While our χViT demonstrates strong results compared to currently available remote sensing specific models, none of them outperforms general-purpose vision models like DINOv2. These findings highlight the necessity of benchmarks like GeoCrossBench to advance robust foundation models for comprehensive Earth observation.

1. Introduction

The growth of remote sensing data and satellite imagery in particular (Gorelick et al., 2017; Zhu et al., 2017; Ma et al., 2019) has led to the development of sophisticated deep learning models capable of analyzing complex geospatial patterns and dynamics. Among these, pre-trained foundation models have emerged as a popular paradigm for learning generalizable representations from vast and diverse remote sensing (RS) datasets (Xiong et al., 2024; Fuller et al., 2023; Cong et al., 2022; Han et al., 2024; Tseng et al., 2025; Jakubik et al., 2023; Wang et al., 2024). Such RS data is inherently multimodal, with sensors capturing information across various *bands* of the electromagnetic spectrum, including multispectral, hyperspectral, and synthetic aperture radar (SAR) (Torres et al., 2012; Drusch et al., 2012; Roy et al., 2014; Guanter et al., 2015). While these foundation models have shown transfer on downstream tasks for the same sensors and bands, their transfer to inputs with different sensors and bands, their **cross-band generalization**, is an issue for practical applications. This type of generalization determines how well a model transfers between different spectra and modalities such as from RGB optical to near-infrared (NIR) or SAR. Robust generalization across spectral domains is crucial for creating more versatile and practical RS models.

We introduce **GeoCrossBench** to assess cross-band generalization in remote sensing and guide progress. GeoCrossBench focuses on three canonical remote sensing tasks: scene classification, semantic segmentation, and change detection, covering both Sentinel-2 optical/multispectral data and Sentinel-1 SAR data. Specifically, we build GeoCrossBench from the GeoBench datasets (Lacoste et al., 2023) and enrich them with additional public datasets that widen the range of resolutions and geographic contexts. For the datasets missing SAR bands we fuse the Sentinel-2 multispectral bands (RGB, NIR, ...) with co-registered Sentinel-1 SAR bands (VV/VH dual-polarization). The core idea of GeoCrossBench is to train models on a common band configuration (e.g., RGB) and then evaluate on a variety of unseen bands from both optical and SAR modalities, as illustrated in Figure 1. To provide a comprehensive analysis, we evaluate generalization using two primary settings: full fine-tuning and fine-tuning with frozen backbone. For practical utility and simplicity, models are trained on RGB imagery, given its greater abundance and the more common expertise associated with it over multispectral or radar data.

We evaluate a range of existing and recent foundation models using GeoCrossBench. Building on ChannelViT (Bao et al., 2024), an extension of the Vision Transformer (ViT)

---

[1]YerevaNN, Yerevan, Armenia [2]Yerevan State University, Yerevan, Armenia [3]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada [4]Vector Institute, Toronto, ON, Canada. Correspondence to: Hakob Tamazyan <hakob@yerevann.com>.

Proceedings of TerraBytes: Towards global datasets and models for Earth Observation Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

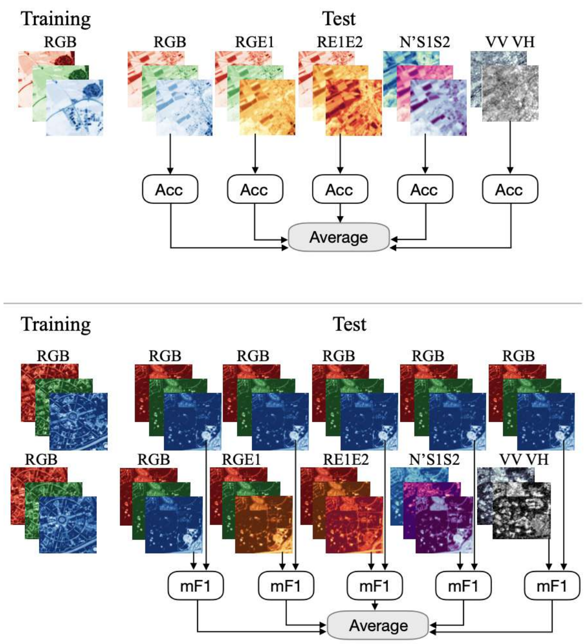


Figure 1. The GeoCrossBench evaluation framework. Models are trained on RGB bands and tested on their generalization to various triplets of optical bands (e.g., RGB, RGE1, RE1E2, N'S1S2) and SAR bands (VV, VH). We report performance for each combination and on average.

(Dosovitskiy et al., 2021) for channel-wise modeling, we develop a new baseline for band-wise modeling in RS. We call this model χ**ViT** (ChiViT), short for **Ch**annel-based **i**BOT pre-trained **ViT**, and pretrain it using the iBOT (Zhou et al., 2022) paradigm on our own large-scale, multi-modal dataset.

We find that many current foundation models struggle with cross-band generalization. Furthermore we discover that RS-specific foundation models fail to outerperform general-purpose vision models like DINOv2 (Oquab et al., 2023). Finally we show that our extension of ChannelViT to remote sensing does in fact deliver improved cross-band transfer to achieve competitive results. These experiments underscore the pressing need for a rigorous and standardized benchmarks like GeoCrossBench.

On publication we will share the GeoCrossBench data, code, and models. This full release can help measure progress, identify weaknesses in current approaches, and drive the development of more robust and reliable foundation models.

2. GeoCrossBench Benchmark: Dataset and Evaluation Protocol

GeoCrossBench is designed to rigorously assess the cross-band generalization of remote sensing (RS) foundation models. The primary goal is to evaluate how well models, trained on standard RGB imagery, adapt to unseen spectral band combinations, including those from different sensor types like multispectral optical and Synthetic Aperture Radar (SAR). The benchmark is built upon key RS tasks—scene classification, semantic segmentation, and change detection—utilizing diverse spectral modalities and a standardized evaluation protocol to ensure fair comparisons.

2.1. Datasets

GeoCrossBench extends several datasets from the original GeoBench benchmark by systematically fusing them with Sentinel-1 SAR data where previously absent, and also incorporates new relevant datasets. All datasets within GeoCrossBench leverage 10-band optical data from Sentinel-2 (bands with $\leq$20m resolution) and dual-polarization SAR data from Sentinel-1 (VV, VH), resulting in a consistent 12-band input structure per sample. This includes datasets for scene classification (x-bigearthnet, x-so2sat, x-brick-kiln, x-eurosat), semantic segmentation (x-cashew-plantation, x-SA-crop-type, x-harvey-building, x-sen1floods11), and change detection (x-harvey-flood, x-oscd). The process of integrating Sentinel-1 data involved co-registering SAR imagery with existing optical data, ensuring temporal and spatial alignment suitable for cross-modal learning and evaluation.

2.2. Evaluation Protocol

The core evaluation principle of GeoCrossBench is to train (or fine-tune) models exclusively on the RGB bands (B4, B3, B2 from Sentinel-2) of the training split for each downstream task. Subsequently, these models are evaluated on the test split using a predefined set of 3-channel optical and 2-channel SAR band combinations:

- **RGB:** Sentinel-2 B2, B3, B4 (seen during training).
- **RGE1:** Sentinel-2 B4, B3, B5.
- **RE1E2:** Sentinel-2 B5, B6, B7.
- **N'S1S2:** Sentinel-2 B8A, B11, B12.
- **VV-VH:** Sentinel-1 VV and VH polarizations.

Performance on these unseen band combinations, relative to the in-distribution RGB performance (which is also measured), quantifies the model's cross-band generalization.

For **scene classification**, models assign labels to image patches, with performance (F1Score or Accuracy, depending on the dataset) averaged across the five band settings. For **semantic segmentation**, pixel-level classification quality (mIOU or bIOU) is averaged across these band combinations. For **change detection**, models identify differences between image pairs. During testing, the 'after' image uses the various target band combinations while the 'before' image remains RGB; the final score (bIOU or F1Score) is an average across these conditions. The overall GeoCrossBench score for a model on a task is this averaged metric.

3. Model Comparisons and Baselines

We considered a wide variety of models and fine-tuned them in two primary settings: (i) **full fine-tuning**, where all parameters of the pretrained foundation model and the task-specific head are updated; and (ii) **fine-tuning with frozen backbone**, where only the parameters of a newly added task-specific head (e.g., a linear layer for classification, a decoder for segmentation/change detection) are trained. These settings represent a trade-off between model's training capacity and preservation of the generalization capabilities that might come from pretraining.

3.1. Pre-trained Foundation Models and Supervised Models

Specialized Remote Sensing Foundation Models. We picked most publicly available models pre-trained on remote sensing data having less than 100M parameters (ViT-B and Swin-B), namely ChannelViT (Bao et al., 2024), DOFA (Xiong et al., 2024), SatlasNet (Bastani et al., 2023), CROMA (Fuller et al., 2023), AnySat (Astruc et al., 2024) and Prithvi (Jakubik et al., 2023).

General-purpose Image Foundation Models. We also added several general-purpose models as baselines. We took self-supervised models of self-distillation type iBOT (Zhou et al., 2022), which is pretrained on ImageNet, and DINOv2 (Oquab et al., 2023) pretrained on a huge custom dataset of 145M images. Note that we use ViT-B version of DINOv2 which is actually distilled from a much larger ViT-g teacher. Following (Lacoste et al., 2023), we also fine-tuned ImageNet-pretrained ResNet-50 and ViT-B that have never gone through self-supervised training. Recent work (Xu et al., 2025) has demonstrated that even non-pretrained models can produce competitive results with enough hyperparameter tuning budget. We omitted such baselines as we prefer fine-tuning recipes that are relatively easy and quick to implement for each new downstream task.

3.2. A New Baseline: Self-supervised Channel-ViT on Remote Sensing Data

The ability to learn transferable representations from diverse partially observed spectral inputs is essential for robust cross–band generalization. Motivated by recent advances in multi-channel self-supervision we extend ChannelViT (Bao et al., 2024) with a hierarchical pre-training recipe tailored to remote sensing imagery that we name χViT (ChiViT). The core idea is to give each spectral band equal importance during pretraining such that the network can be a) fine-tuned on any subset of bands available without architectural changes and b) able to exchange information between spectrally distinct modalities.

Pretraining dataset. To pretrain χViT for strong cross-band generalization, we extended Satlas Pretrain dataset (Bastani et al., 2023) up to over 23 million images. This dataset was collected to expose the model to a wide spectrum of Earth's surface characteristics, captured by various spectral bands and resolutions. Notably we added "parallel" data: the BigEarthNet (Sumbul et al., 2021) and Sen12MS datasets (Schmitt et al., 2019), offer Sentinel-1 and Sentinel-2 image pairs that are lined up, crucial for learning joint radar-optical features.

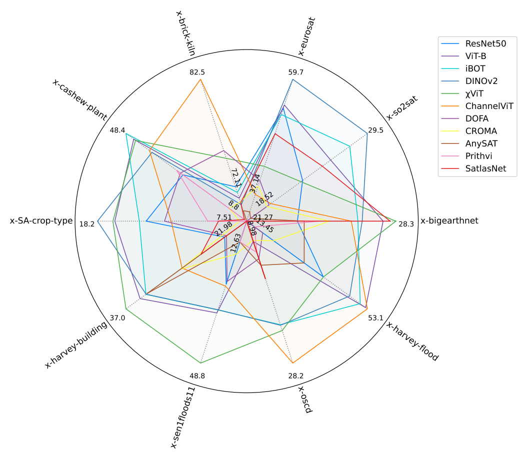
4. Results

Experiments on GeoCrossBench highlight several key findings. Firstly, remote sensing foundation models demonstrate constrained capabilities in generalizing across spectral bands. This limitation is particularly pronounced when dealing with bands that diverge significantly from standard RGB, such as Synthetic Aperture Radar (SAR), suggesting that current RS-specific pretraining may not adequately capture the nuances of diverse spectral data. Secondly, general-purpose image foundation models, notably DINOv2 (especially its frozen backbone version), achieve surprisingly robust cross-band performance, often outperforming foundation models specifically pretrained for RS. Thirdly, while full fine-tuning generally leads to better accuracy, there are instances, where a frozen backbone excels, as seen in Figure 3. All these findings are supported by model rankings in Table 1 and visual summaries in Figure 2, where Figure 2a

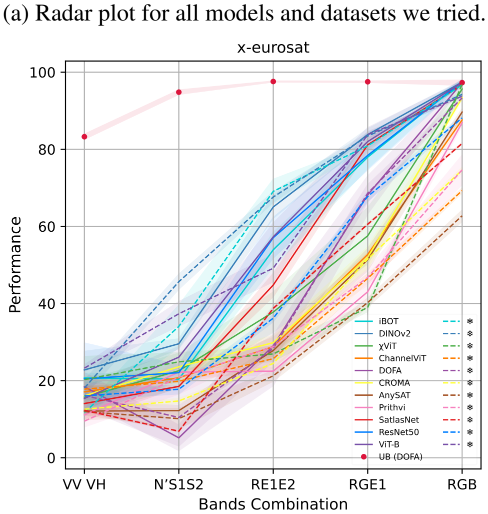
Table 1. Per-task ranking of all tested models on GeoCrossBench. The last column indicates the average rank across 10 tasks.

	x-ben	x-so2sat	x-eurosat	x-cashew-plant	x-SA-crop	x-brick-kiln	x-harvey-segm	x-harvey-change	x-sen1floods11	x-oscd	Average
DINOv2❄	1	1	1	7	2	2	4	5	6	7	**3.60**
χViT	4	8	8	3	3	5	1	6	1	4	**4.30**
iBOT❄	2	4	3	6	5	6	3	8	7	1	**4.50**
iBOT	8	3	6	1	6	10	5	3	4	6	**5.20**
DINOv2	7	2	2	5	1	15	6	4	5	5	**5.20**
ViT-B	6	5	4	2	4	11	2	2	3	17	**5.60**
ChannelViT	9	11	13	4	9	1	10	1	8	3	**6.90**
χViT❄	3	10	12	12	11	4	8	11	2	12	**8.50**
ResNet50	14	7	5	10	7	9	13	7	9	20	**10.10**
DOFA	16	13	9	9	8	3	15	17	10	14	**11.40**
DOFA❄	15	12	10	15	10	7	14	14	11	9	**11.70**
SatlasNet	5	6	7	18	16	13	11	19	19	10	**12.40**
ChannelViT❄	10	15	18	13	13	20	12	10	17	2	**13.00**
CROMA	11	14	11	16	19	12	9	15	12	18	**13.70**
AnySat	13	17	15	19	20	16	7	9	13	13	**14.20**
SatlasNet❄	19	9	14	17	12	14	17	12	20	8	**14.20**
Prithvi	12	18	16	8	14	18	19	18	14	15	**15.20**
CROMA❄	17	16	19	20	17	8	18	13	16	16	**16.00**
AnySat❄	20	20	20	11	18	19	16	16	15	11	**16.60**
Prithvi❄	18	19	17	14	15	17	20	20	18	19	**17.70**

shows metrics for each model and task averaged across band combinations and Figure 2b shows the drop of the performance of all models on the band combinations of x-eurosat dataset. The results for the fine-tuned DOFA on the test bands, shown in this second figure, serve as an upper bound for its generalization from RGB to these band combinations.



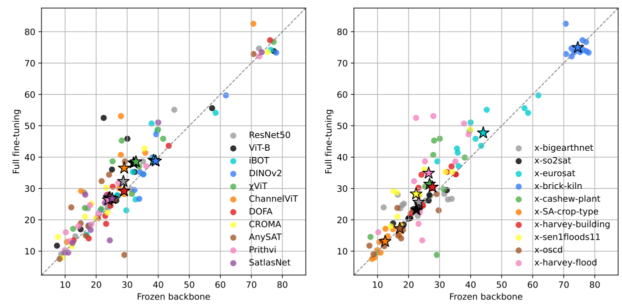
(a) Radar plot for all models and datasets we tried.



(b) Detailed results for x-eurosat dataset.

Figure 2. Quick summary of the main results on GeoCrossBench.

The experimental results prompt discussion on several critical points. The overall value of current RS-specific pretraining methods is brought into question, as general-purpose models leveraging common visual features (e.g., shapes, contours) perform competitively. This suggests a potential need for novel pretraining strategies designed to capture deeper, more fundamental relationships between different spectral bands. Regarding the possibility of benchmark saturation, our analysis indicates that GeoCrossBench has not reached its limits. Non-RGB bands contain substantial information relevant to the tasks, implying that there is considerable room for improvement in model performance.



(a) Colored by models (b) Colored by datasets

Figure 3. Performance of models with frozen backbone (x-axis) vs. full fine-tuning (y-axis). Each point is a model-dataset pair.

5. Related Work

Remote sensing presents unique data challenges, motivating specialized foundation models (Rolf et al., 2024) to handle vast unlabeled data and enable efficient transfer learning. Pioneering works like SatMAE (Cong et al., 2022) and Satlas (Bastani et al., 2023) established self-supervised and large-scale supervised pre-training for RS, respectively, while Scale-MAE (Reed et al., 2023) focused on generalization across spatial resolutions. Our work, GeoCrossBench, complements these by addressing the critical challenge of *spectral* generalization across bands. While many RS foundation models learn from varied spectral bands, employing techniques like auto-encoding (e.g., SatMAE (Cong et al., 2022), MMEarth (Nedungadi et al., 2024)), learning separate intra-modal representations (e.g., SoftCon (Wang et al., 2024), DOFA (Xiong et al., 2024)), or joint inter-modal representations (e.g., CROMA (Fuller et al., 2023), AnySat (Astruc et al., 2024), Galileo (Tseng et al., 2025)), they typically do not focus on generalization to unseen bands. GeoCrossBench targets this crucial gap, highlighting the need for models to adapt across spectral inputs.

Conclusion

We introduce GeoCrossBench to evaluate cross-band generalization in remote sensing foundation models. Our experiments reveal that current remote sensing-specific foundation models do not yet significantly outperform general-purpose vision models in cross-band tasks. We hope that GeoCrossBench will serve as a catalyst for future research.

Acknowledgements

The research was supported by the Higher Education and Science Committee of MESCS RA (Research project No 24RL-1B049). This work was also supported by the Strategic Armenian Science and Technology Investment Community (SASTIC).

References

Astruc, G., Gonthier, N., Mallet, C., and Landrieu, L. Anysat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024.

Bao, Y., Sivanandan, S., and Karaletsos, T. Channel vision transformers: An image is worth 1 x 16 x 16 words. 2024.

Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., and Kembhavi, A. 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.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. volume 35, pp. 197–211, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Bargellini, P., and Latorre, C. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, 120:25–36, 2012.

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

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Mielke, C., Erhard, M., Boesche, N., Mihailova, B., Schrader, J., Leitão, P. J., Rabe, A., Doerffer, R., Fischer, S., Stuffler, T., Müller, R., Richter, R., Oppermann, K., Bachmann, M., Müller, A., Sang, B., Walter, I., Bieniarz, J., Brosinsky, A., Eckardt, A., Heiden, U., Heldens, W., Hill, J., Hostert, P., Krasemann, H., van der Linden, S., Mauser, W., Oppelt, N., Roettgers, R., Schneiderhan, T., Staenz, K., and Wulf, H. The enmap spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, 7(7):8830–8857, 2015.

Han, B., Zhang, S., Shi, X., and Reichstein, M. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27852–27862, 2024.

Jakubik, J., Roy, S., Phillips, C., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., and Lang, N. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. 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.

Rolf, E., Klemmer, K., Robinson, C., and Kerner, H. Mission critical–satellite data is a distinct modality in machine learning. In *ICML (Position Papers)*, 2024.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, J., Larsen, T. A. K., Masek, J. G., Schott, J. R., Radeloff, V. C., Schaaf, C. L. S., Cohen, W. B., Crawford, C. J., Helmer, E. T., Goward, S. N., O'Connell, P. E., Williams, D. L., Shuchman, R. A., and Nemani, R. R. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

Schmitt, M., Hughes, L. H., Qiu, C., and Zhu, X. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:153–160, 2019.

Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., and Markl, V. 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.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas Traver, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., and Rostan, F. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120: 9–24, 2012.

Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J. R., Shelhamer, E., Kerner, H., and Rolnick, D. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025.

Wang, Y., Albrecht, C. M., and Zhu, X. X. Multi-label guided soft contrastive learning for efficient earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Xiong, Z., Wang, Y., Zhang, F., Stewart, A. J., Hanna, J., Borth, D., Papoutsis, I., Saux, B. L., Camps-Valls, G., and Zhu, X. X. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Xu, Z., Gupta, R., Cheng, W., Shen, A., Shen, J., Talwalkar, A., and Khodak, M. Specialized foundation models struggle to beat supervised baselines. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A. L., and Kong, T. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.