GEOCROSSBENCH: CROSS-BAND GENERALIZATION FOR REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

The number and diversity of remote sensing satellites grows over time, while the vast majority of labeled data comes from older satellites. As the foundation models for Earth observation scale up, the cost of (re-)training to support new satellites grows too, so the generalization capabilities of the models towards new satellites become increasingly important. In this work we introduce GeoCrossBench, an extension of the popular GeoBench benchmark with a new evaluation protocol: it tests the in-distribution performance; generalization to satellites with no band overlap; and generalization to satellites with additional bands with respect to the training set. We also develop a self-supervised extension of ChannelViT, χViT, to improve its cross-satellite performance. First, we show that even the best foundation models for remote sensing (DOFA, TerraFM) do not outperform general purpose models like DINOv3 in the in-distribution setting. Second, when generalizing to new satellites with no band overlap, all models suffer 2-4x drop in performance, and χViT significantly outperforms the runner-up DINOv3. Third, the performance of all tested models drops on average by 5-25% when given additional bands during test time. Finally, we show that fine-tuning just the last linear layer of these models using oracle labels from all bands can get relatively consistent performance across all satellites, highlighting that the benchmark is far from being saturated. We publicly release the code and the datasets to encourage the development of more future-proof remote sensing models with stronger cross-satellite generalization.

1 INTRODUCTION: GENERALIZATION ACROSS REMOTE SENSING DATA

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; Danish et al., 2025; Jakubik et al., 2023; Wang et al., 2024b). 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). These models promise ease-of-use and transfer across RS data.

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 and costly to achieve by retraining. This type of generalization determines how well a model transfers between different spectra and modalities such as from RGB optical to SAR.

This is a critical gap: real-world applications can require models to summarize data from various sensors, to adapt to new spectral bands as sensor technology evolves, or to do a new task that needs bands complementary to the training bands. Robust generalization across spectral domains is crucial for creating more versatile and pracical remote sensing models, because large-scale training and fine-tuning is not accessible for all researchers and practitioners.

We introduce **GeoCrossBench** to assess the gap of cross-band generalization in remote sensing with *three* complementary evaluation protocols: (1) *in-distribution* – train and test on the same bands, (2)

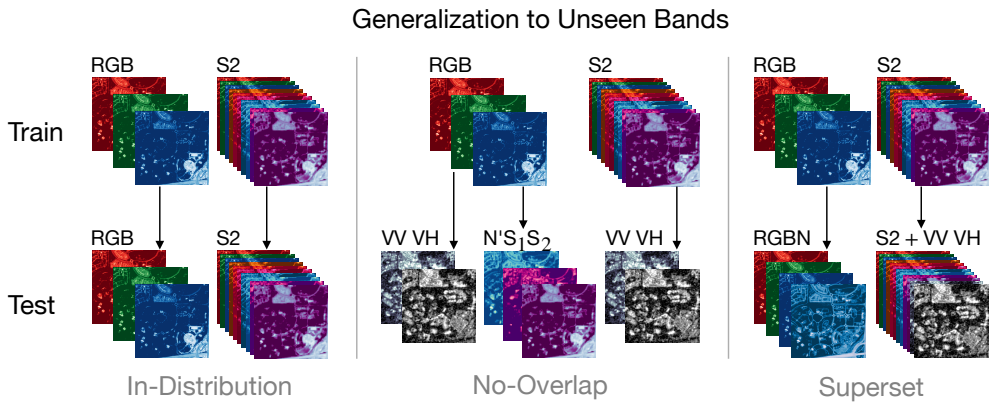


Figure 1: The GeoCrossBench evaluation framework. (1) *In-Distribution*: fine-tune on RGB and evaluate on RGB; fine-tune on full S2 and evaluate on S2. (2) *No-Overlap*: evaluate transfer from RGB→S1 (VV, VH), RGB→N'S_1S_2 (B8A, B11, B12) and S2→S1. (3) *Superset*: RGB→RGBN (RGB+NIR) and S2→S2+S1 (optical+SAR fusion).

no overlap bands generalization – train on optical bands and test on not overlapping bands, and (3) *superset bands generalization* – test-time inputs provide strictly more bands than used in training.

GeoCrossBench focuses on three canonical remote sensing tasks: scene classification, semantic segmentation, and change detection, covering both Sentinel-2 (S2) optical/multispectral data and Sentinel-1 (S1) 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. Moreover, for the datasets missing SAR bands we fuse the Sentinel-2 multispectral bands with co-registered Sentinel-1 SAR bands (VV/VH dual-polarization). This fusion expands the spectral range of the datasets to allow for more rigorous cross-band evaluation. The core idea of GeoCrossBench is to train models on a common band configuration (e.g., RGB, S2) 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 that also considers practical computational constraints, we evaluate generalization using two primary settings: full fine-tuning and fine-tuning with frozen backbone.

We systematically 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) (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.

Experiments with our benchmark reveal insights into current performance and potential directions of improvement. We find that many foundation models struggle with cross-band generalization. Furthermore, we discover that RS-specific foundation models fail to outerperform general-purpose vision models like DINOv3 Siméoni et al. (2025) in the in-distribution setting. Finally we show that χViT model delivers improved cross-band transfer and achieves best results under these settings. Findings 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 ultimately drive the development of more robust, versatile, and reliable foundation models for comprehensive Earth observation.

2 GeoCrossBench Benchmark: Dataset and Evaluation Protocol

GeoCrossBench is designed to thoroughly evaluate the ability of remote sensing foundation models to generalize knowledge learned from one set of spectral bands (specifically RGB or S2) to other spectral band combinations, that either match, do not overlap, or strictly supersets the training bands.

2.1 Motivation and Design Principles

The primary motivation behind GeoCrossBench is the observation that many foundation models, despite achieving high performance on tasks when training and testing data come from the same spectral distribution, but their performance often degrades when processing imagery with different spectral characteristics. This limitation hinders their practical utility. This generalization challange is not a theoretical concern, it poses a significant barrier to deploying models in a constantly evolving satellite ecosystem where large-scale labeled datasets are often unavailable for newer commercial satellites. A practitioner might need to transfer a model trained on public Sentinel-2 data to newer platforms like Planet SuperDove or Satlantis GARAI, which share some spectral bands with Sentinel-2 but also introduce new ones (e.g., Green I, Yellow). More extreme generalization is required when transferring between entirely different sensor types. For example, adapting an optical model from Sentinel-2 for use with SatVu's HotSat, which captures purely thermal data, requires generalization to a non-overlapping spectral range. The same challenge arises when transferring from optical to SAR imagery. While no benchmark can perfectly replicate every possible transfer task, GeoCrossBench provides a standardized proxy for these real-world challenges, using the best available large-scale labeled data to foster the development of more robust and future-proof foundation models.

GeoCrossBench is built on the following principles:

- **Focus on Generalization:** The main goal is to evaluate how well models adapt from a seen spectral inputs to unseen ones.
- **RS Specific Tasks:** Evaluation is based on tasks central to remote sensing: scene classification, semantic segmentation, and change detection.
- **Diverse Spectral Modalities:** The benchmark incorporates both multi-band optical data and dual-polarization SAR data to test generalization across fundamentally different sensing mechanisms.

2.2 Datasets

GeoCrossBench extends the original GeoBench benchmark by fusing them with corresponding Sentinel-1 SAR data and also incorporates completely new datasets relevant to cross-band generalization, such as x-sen1floods11, x-oscd, x-harvey-flood and x-harvey-building. All datasets in GeoCrossBench utilize Sentinel-2 10-band optical data (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12 – bands with $\leq$20m resolution) and Sentinel-1 dual-polarization SAR data (VV, VH – absolute values of the complex numbers), resulting in a 12-band input for each sample. An overview of the datasets is provided in Table 1. The difference between the x-harvey dataset and the original (Rudner et al., 2019) lies in the split we provide, which bypasses the geographical distribution shift. Additionally, we use the original dataset to construct a change detection task by pairing pre- and post-flood images, along with the corresponding flood segmentation masks. x-sen1floods11 is a subset of the original Sen1Floods11 dataset (Bonafilia et al., 2020b), created by removing the weakly labeled portion.

Bringing Sentinel-1 data. For the OSCD dataset (Caye Daudt et al., 2018), we combined it with the corresponding Sentinel-1 data collected by another team (Hafner et al., 2022) to create x-oscd. The m-so2sat dataset (Lacoste et al., 2023; Zhu et al., 2020) from GeoBench already includes paired Sentinel-1 bands. We apply the following transformation to obtain absolute values: $vh = 10 \cdot \log_{10}(vh_i^2 + vh_r^2 + \varepsilon)$, where $\varepsilon = 10^{-10}$, and create x-so2sat. We apply the same transformation to obtain the vv band. For m-eurosat (Lacoste et al., 2023), we retrieve the corresponding Sentinel-1 data from EuroSAT-SAR (Wang et al., 2025) to create x-eurosat. We create x-bigearthnet by pairing m-bigearthnet images (Lacoste et al., 2023) with those from the original set (Sumbul et al., 2021) whose Sentinel-2 parts match those in GeoBench, and then retrieve the corresponding Sentinel-1 images. We create x-cashew-plantation by pairing m-cashew-plantation (Lacoste et al., 2023) with the corresponding Sentinel-1 images retrieved from the Copernicus Open Access Hub (European Space Agency, 2025) using the dates provided in the Sentinel-2 version of the original data. The m-brick-kiln dataset (Lacoste et al., 2023; Lee et al., 2021) does not contain temporal extent information for the imagery. To address this, we collected all available cloud-free Sentinel-2 acquisitions between October 2018 and May 2019. For each sample in m-brick-kiln, we selected a pixel-level similar image from our collected data, recorded its acquisition date, and retrieved the corresponding Sentinel-1 image from the Copernicus Open Access Hub. Using this approach, we constructed x-brick-kiln. We created x-SA-crop-type from the m-SA-crop-type (Lacoste et al., 2023). The original set (Western

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

Cape Department of Agriculture and Radiant Earth Foundation, 2021) contains temporally close Sentinel-1 and Sentinel-2 image pairs, where each Sentinel-1 image was selected as the closest available in time to its corresponding Sentinel-2 image. In m-SA-crop-type, 100 Sentinel-2 images were rotated. To establish accurate matches between Sentinel-1 and Sentinel-2 images, we replaced these rotated images with nearest images from the original dataset.

2.3 Evaluation Protocol

GeoCrossBench evaluates models under three distinct settings designed to probe different aspects of generalization. For all settings, models are fine-tuned on the training data and then evaluated on the test data from various downstream datasets, each representing one of the three remote sensing tasks.

Setting 1: In-Distribution. This setting establishes a baseline performance metric. Models are trained and evaluated on the same set of bands. This measures the model's effectiveness in a standard, non-generalization setting. We test two common configurations:

- **Train on RGB → Evaluate on RGB:** Models are fine-tuned using only Sentinel-2's RGB bands (B4, B3, B2).
- **Train on S2 → Evaluate on S2:** Models are fine-tuned using all 10 available Sentinel-2 bands.

Setting 2: No-Overlap Bands. This setting tests a model's ability to transfer learned representations to a completely different sensor type, representing a challenging zero-shot generalization task.

- **Train on RGB → Evaluate on S1:** Generalization from RGB to SAR.
- **Train on S2 → Evaluate on S1:** Generalization from multispectral optical to SAR.
- **Train on RGB → Evaluate on N'S_1S_2:** Generalization from RGB to narrow near infrared, shortwave infrared 1 and 2 bands (S2 B8A, B11, B12).

Setting 3: Superset Bands. This setting assesses a model's robustness and ability to leverage new information when presented with more spectral bands at test time than it was trained on. This simulates a real-world scenario where a model trained on legacy data must operate on data from a newer, more capable sensor.

- **Train on RGB → Evaluate on RGBN:** Tests generalization from 3-band optical to 4-band optical, adding the Near-Infrared band (S2 B8).
- **Train on S2 → Evaluate on S2+S1:** Tests generalization from 10-band multispectral to a 12-band fused optical-SAR product.

Scene Classification. For scene classification tasks, models are trained to assign a class label to an entire image patch. The testing is done by using the evaluation band combination as input and the performance is measured using the corresponding evaluation metric for each task: **F1Score** for x-bigearthnet, **Accuracy** for x-so2sat, x-eurosat and x-brick-kiln.

Semantic Segmentation. For semantic segmentation, models are trained to assign a class label to each pixel in an image. At test time, the model segments images using the corresponding band combinations. Segmentation quality is measured by: **mIOU** for x-cashew-plantation, x-SA-crop-type, and x-sen1floods11, and **bIOU** for x-harvey-building.

Change Detection. Change detection tasks require the model to identify differences between two remote sensing images of the same area taken at different times. The evaluation involves training on image pairs of the same band combinations and testing on pairs where the 'after' image is replaced with different band combinations, while the 'before' image bands will be kept the same. We report **bIOU** for x-harvey-flood and **F1Score** for x-oscd.

3 Comparisons: Foundation Models, Supervised Models, and a New Baseline

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 TerraFM Danish et al. (2025), 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). The details of adapting the bands we need in the benchmarks to the expected inputs of the models are in Appendix E.

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, DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) models pretrained on a huge custom dataset of images. Note, as we are using models having less than 100M parameters, only ViT-B version of DINOv2 and DINOv3 models are used. Moreover, models like the satellite version of DINOv3 (Siméoni et al., 2025) or DINO-MC (Wanyan et al., 2024) lack the ViT-B version and they are not included in our comparison. 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.

Hyperparameters. For the original GeoBench tasks, we used a fixed set of hyperparameters selected from prior related works that report their hyperparameters for specific tasks. For segmentation and change detection tasks, we use the UperNet head for all models, which takes the internal representations from 4 layers of the encoder. For change detection, we compute the difference between the encoder representations of the two input images. For the other four tasks, we used a quick hyperparameter search. Refer to Appendix E.1 for the details. Input sizes are chosen to match the size used during pretraining of the underlying model (following (Corley et al., 2024)).

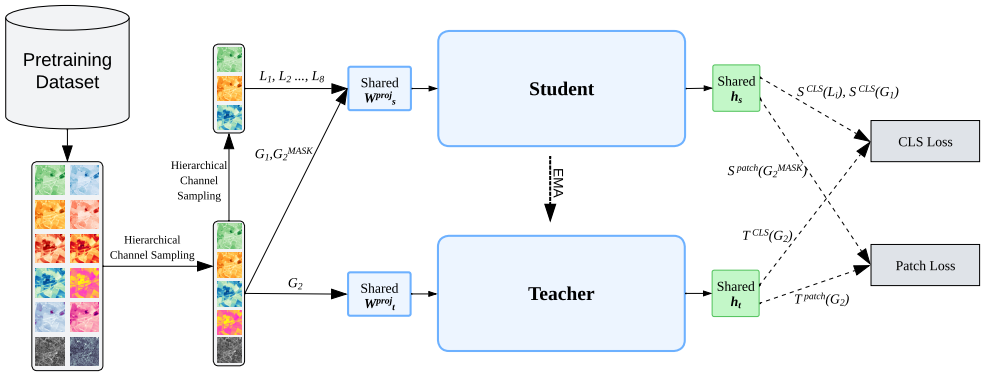


Figure 2: Overview of the iBOT-style self-distillation pretraining used for χViT. Hierarchical channel sampling is applied to create distinct views for the student and teacher, where student channels are a subset of the teacher's channels. Shared projection weights and a shared prediction head are utilized, with losses computed for both CLS and patch tokens.

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.

Architecture. ChannelViT preserves channel-specific information by tokenizing each single band independently and adding a learnable *channel embedding* $\mathbf{e}^{\text{chn}}_c$ that is analogous to the positional embedding $\mathbf{e}^{\text{pos}}(h, w)$ of ViT. Given an input $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, we partition every band into $N_p = (H/P)\cdot(W/P)$ non-overlapping patches of size $P\times P$. Unlike standard ViTs that create a single token from a multi-channel patch, ChannelViT generates one token from each single-channel patch. Thus, for each channel $c \in \{1, ..., C\}$ and each spatial patch $j \in \{1, ..., N_p\}$, we obtain a patch $x_{c,j}$. Each such single-channel patch is flattened into a vector of dimension P^2. These flattened patches are then linearly projected into D-dimensional embeddings using a learnable linear projection $W \in \mathbb{R}^{P^2\times D}$. Crucially, these projection weights W (image filters) are shared across all channels, promoting the learning of shared low-level features and enhancing robustness (Bao et al., 2024). To retain spatial and channel-specific information, learnable positional embeddings $\mathbf{e}^{\text{pos}}_j \in \mathbb{R}^D$ (shared across channels) and learnable channel embeddings $\mathbf{e}^{\text{chn}}_c \in \mathbb{R}^D$ are added to each projected patch token. A learnable classification token, $\mathbf{e}^{\text{CLS}} \in \mathbb{R}^D$, is prepended to the sequence. The final sequence of $N = C \cdot N_p + 1$ tokens fed to the Transformer encoder is structured as: $[\mathbf{e}^{\text{CLS}}; \ldots; Wx_{c,j} + \mathbf{e}^{\text{pos}}_j + \mathbf{e}^{\text{chn}}_c; \ldots]$. This allows the model to reason across both spatial locations and spectral channels simultaneously.

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. Please refer to Appendix C for all details.

We performed a small search over certain details of the training algorithm that are reported in Appendix D.

4 GeoCrossBench Experiments and Discussion

Table 2 shows the ranking of all tested models on each evaluation setting. Figure 3 provides a visual summary of key findings.

RS foundation models do not outperform general-purpose models in-distribution. Our first key finding from *In-Distrubution* setting is that even the best foundation models designed specifically for remote sensing, such as DOFA or TerraFM, fail to consistently outperform general-purpose vision models like DINOv3. When trained and tested on the same band combinations (e.g., RGB $\rightarrow$ RGB or S2 $\rightarrow$ S2), DINOv3 achieves competitive, and in several cases superior, performance. This suggests that the large-scale, diverse pre-training of general-purpose models provide a powerful and transferable feature foundation that is not yet surpassed by domain-specific pre-training on RS data alone.

RS foundation models are limited in their cross-band generalization. The limitations of current foundation models become apparent when generalizing to unseen bands. Under *No-Overlap* setting, which tests generalization to satellites with no band overlap (e.g., S2 $\rightarrow$ S1), all models suffer a severe 2-4x drop in performance. This highlights a fundamental weakness in transferring learned knowledge across different sensor modalities. However, within this challenging setting, our proposed model, χViT, significantly outperforms all other contenders, including the strong runner-up DINOv3.

This weakness is further confirmed by *Superset* setting. Counter-intuitively, providing models with more information at test time by including additional bands also leads to a performance drop, with models degrading by 5-25% on average. This suggests that current architectures may overfit to the specific number and distribution of input channels, failing to robustly integrate novel spectral information without explicit fine-tuning.

Fine-tuning is often necessary for adequate accuracy. On average, all models with full fine-tuning outerperform their frozen counterparts. Refer to Appendix B for detailed analysis. There are a few exceptions if we consider only *No-Overlap* scenario. Namely, DINOv2 and TerraFM are slightly better with frozen backbones (Table 2).

The value of RS-specific pretraining. RS-specific pretraining is not delivering top performance on GeoCrossBench against pretraining methods for regular RGB imagery. First, this raises a question:

Table 2: Performance evaluation of all tested models on GeoCrossBench. The last column indicates the average score across all our settings.

	In-Distribution				No-Overlap					Superset				
Fine-tuned on	**RGB**	**S2**	**AVG**	**#**	**RGB**	**S2**	**RGB**	**AVG**	**#**	**RGB**	**S2**	**AVG**	**#**	**Overall**
Tested on	**RGB**	**S2**			**S1**	**S1**	**N'S_1S_2**			**RGBN**	**S2+S1**			**AVG**
χViT	61.81	63.53	62.67	6	17.96	20.93	30.37	23.09	1	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	3	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	6	52.36	57.04	54.7	3	42.32
ViT-B	62.77	62.72	62.75	4	18.87	14.75	25.18	19.6	8	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	9	54.52	47.59	51.06	5	41.16
χViT*	56.95	58.42	57.69	12	19.02	18.92	27.12	21.69	2	47.75	48.6	48.17	8	39.54
DINOv2*	61.77	56.58	59.17	9	16.28	15.84	30.13	20.75	5	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	21	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	12	40.76	50.23	45.5	9	37.92
SatlasNet	49.23	69.12	59.18	8	14.62	14.61	15.4	14.88	16	38.88	58.58	48.73	7	37.21
iBOT*	62.02	50.9	56.46	14	15.3	13.94	30.08	19.78	7	48.63	38.13	43.38	12	37.0
ResNet50	60.43	57.35	58.89	10	13.29	13.48	19.69	15.48	14	41.53	48.33	44.93	11	36.3
DINOv3*	58.51	49.0	53.75	17	14.91	17.17	30.37	20.81	4	46.69	33.54	40.11	17	35.74
DOFA*	59.1	58.03	58.57	11	15.23	14.46	14.83	14.84	17	38.77	45.05	41.91	15	35.07
TerraFM*	56.7	56.72	56.71	13	14.91	12.16	24.81	17.29	10	40.94	35.24	38.09	18	34.5
CROMA	51.58	56.5	54.04	16	16.18	12.26	16.71	15.05	15	34.04	51.61	42.82	14	34.13
ViT-B*	53.42	50.02	51.72	19	16.18	14.03	21.65	17.29	11	42.87	39.84	41.35	16	34.0
Prithvi	52.61	56.88	54.74	15	13.96	11.87	14.71	13.52	22	32.82	53.04	42.93	13	33.7
ResNet50*	53.13	50.22	51.68	20	12.33	13.44	17.81	14.53	19	41.61	31.04	36.32	19	31.37
SatlasNet*	43.6	51.74	47.67	21	12.18	13.5	13.45	13.04	23	24.9	37.17	31.04	20	28.08
CROMA*	40.99	49.7	45.34	22	14.57	13.0	15.6	14.39	20	20.14	37.48	28.81	21	27.35
AnySAT	47.34	58.81	53.08	18	16.19	13.79	18.3	16.09	13	14.72	13.77	14.24	23	26.13
Prithvi*	43.96	28.39	36.17	24	12.68	10.04	13.29	12.0	24	30.69	26.05	28.37	22	23.58
AnySAT*	40.35	47.01	43.68	23	13.35	13.57	17.52	14.81	18	13.0	12.64	12.82	24	22.49

how can these general-purpose models perform transfer at all? One possible explanation is that there are certain correlations between RGB and other bands, especially the features covering the shapes and contours of the objects. The models can learn these patterns from RGB and apply them on other band combinations.

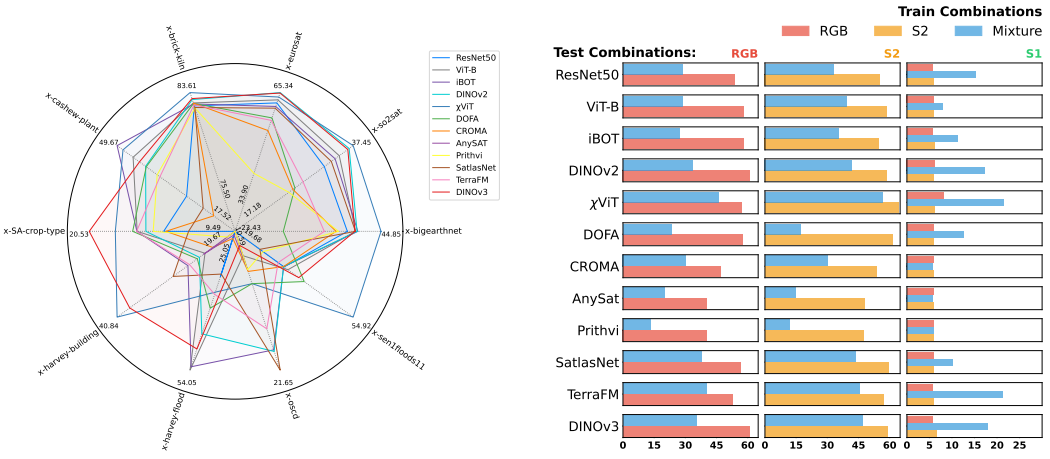
Second, one can ask what additional knowledge can RS-specific foundation models learn that will help them beat general-purpose models. The good performance of χViT on *No-Overlap* setting hints that additional value can come from careful mixing of images from various satellites during pretraining so that the models can learn more complex cross-band relationships between than simple correlations. The poor performance of all tested models in *Superset* setting implies that new ideas are necessary for the models to leverage the additional signal coming from unseen bands at test time. This is a feature that general-purpose models are unlikely to obtain.

Is the benchmark saturated? One way to show that the benchmark is not saturated is to measure the ability of the models to show improved performance when oracle-labeled data is available for other satellites. We examine the potential of the frozen representations of the pretrained models. We perform linear probing on the mixture of representation vectors from four band combinations: RGB, Sentinel-2, Sentinel-1 and N'S_1S_2 using the most challenging classification task: x-so2sat. Then we evaluate these linear models on each of the four combinations. We compare these results with the linear models trained on only RGB, and on only S2. As seen in Fig. 3b, the performance on S1 can be significantly improved for certain backbones (e.g. χViT, DINOv2, DINOv3, TerraFM). This improvement comes with a trade-off: the performance on RGB and S2 slightly decreases with mixture training.

Implications for future models. One of the reasons for the relatively strong performance of χViT compared to other multispectral models might be the trick of *sampling of the bands* during pretraining. The models might learn to rely less on band-specific features and instead focus on patterns shared across bands, which then improves cross-band generalization performance. Sampling of channels during fine-tuning might also be beneficial.

The impact of the *scale of the models and datasets* is relatively underexplored in remote sensing. While usually models based on ViT-L outperform similarly trained models based on ViT-B, the usefulness of scaling RS models towards billions of parameters has yet to be demonstrated. While GeoCrossBench limits the number of parameters during the inference, larger models can still be helpful through distillation, e.g. by using techniques demonstrated in DINOv2.

Finally, the *quality and the quantity of pretraining RS data* can have a significant impact on benchmarks like GeoCrossBench. We expect future work to focus specifically on "parallel" imagery datasets. Just like high quality translation data in the pretraining corpora of LLMs can improve



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

(b) Linear probing results on x-so2sat.

Figure 3: Quick summary of the main results on GeoCrossBench.

knowledge sharing between languages, pairs of images covering the same area from different satellites might improve cross-band generalization abilities of RS foundation models.

5 Related Work

The development of foundation models for Earth observation has seen rapid progress, aiming to create versatile models applicable to a wide array of downstream tasks.

Foundation Models for Remote Sensing. Remote sensing data is in effect its own data modality, with unique challenges and opportunites, and so there has been a call for specialized machine learning approaches and models (Rolf et al., 2024). This large-scale/small-scale tension motivates foundation modeling to enable more efficiency transfer and application across tasks with limited labels. We highlight pioneering methods that established the topic: SatMAE (Cong et al., 2022) demonstrated self-supervised pre-training for RS at ViT scale and Satlas (Bastani et al., 2023) by contrast demonstrated large-scale and multi-task supervised pre-training for RS. Scale-MAE (Reed et al., 2023) refined self-supervised learning for RS by focusing on different spatial resolutions and generalization across them by controlling for ground sample distance (i.e. the physical size of a pixel). Our χViT model and GeoCrossBench dataset is related in spirit to Scale-MAE but spectral, rather than spatial, and explores how to generalize acros bands instead of resolutions. GeoCrossBench is a response to the call to do machine learning for remote sensing: it measures the specific need in RS to generalize across bands given the variety of satellites and the varying coverage of data from each.

Multi-modal/sensor/band Learning in Remote Sensing. Multi-modal data with many and different bands is common in remote sensing due to the existence of multiple satellites. As in self-supervised deep learning for other modalities, foundation models in RS learn from multi-modal data in RS: SatMAE (Cong et al., 2022), Scale-MAE (Reed et al., 2023), and MMEarth (Nedungadi et al., 2024) auto-encode multispectral optical data and MMEarth decodes other modalities; SoftCon (Wang et al., 2024b), DeCUR (Wang et al., 2024a), and DOFA (Xiong et al., 2024) separately learn intra-modal representations of multi-spectral and radar data; and CROMA (Fuller et al., 2023), AnySat (Astruc et al., 2024), TerraFM (Danish et al., 2025) and Galileo (Tseng et al., 2025) jointly learn inter-modal representations of multi-spectral and radar data (CROMA) and more modalities like elevation or climate (AnySat, Galileo). In summary these works explore many ways to learn *from* bands but not *across* bands and do not cover how to extend or generalize to new or different bands. GeoCrossBench highlights this direction of improvement, and measures the need for improvement, which is practically motivated by the cost to (re-)train these ever larger foundation models. It is not feasible to train for all combinations of bands, at least not for most groups, so generalization is necessary.

Datasets and Benchmarking for Remote Sensing. Shared datasets and benchmarks are key for comparability in the context of the diversity of RS data and tasks. Our focus is evaluation, like GEO-Bench (Lacoste et al., 2023), and not pre-training, like Terra (Chen et al., 2024). There are many and high-quality task-specific benchmarks (for marine debris (Kikaki et al., 2024), floods (Bonafilia et al., 2020a), agriculture (Garnot and Landrieu, 2021; Rußwurm et al., 2019; Tseng et al., 2021), and more) but they do not focus on general capacities like generalization or efficiency. GeoCrossBench is needed because no existing benchmark measures our key question of cross-band generalization.

Limitations and Conclusion

This work focuses on evaluating cross-band generalization capabilities of remote sensing foundation models by extending multiple existing datasets with SAR data. The datasets are limited to static objects and scenes and do not cover moving objects for which it is extremely hard to find parallel optical-SAR imagery. Even if the models achieve perfect scores on our benchmark, they might struggle in detecting moving objects on unseen bands. Our experiments showed that RS-specific foundation models still have a lot of room for improvement to significantly outerperform general-purpose models on cross-satellite generalization in both *No-Overlap* and *Superset* scenarios, and we hope GeoCrossBench will motivate further research in this area.

We have used ChatGPT and Gemini to polish the writing in several sections of this paper.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024.

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*, pages 16772–16782, 2023.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020a.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 210–211, 2020b.

R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2018.

Wei Chen, Xixuan Hao, wu yuankai, and Yuxuan Liang. Terra: A multimodal spatio-temporal dataset spanning the earth. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=I0zpivK0A0.

Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020.

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 multispectral satellite imagery. volume 35, pages 197–211, 2022.

Isaac Corley, Caleb Robinson, Rahul Dodhia, Juan M Lavista Ferres, and Peyman Najafirad. Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 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.

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

European Space Agency. Copernicus Open Access Hub. https://scihub.copernicus.eu, 2025. Accessed: 2025-05-16.

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

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

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.

Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Thomas Kuester, André Hollstein, Gerhild Rossner, Christian Mielke, Marco Erhard, Nadin Boesche, Bistra Mihailova, Jörg Schrader, Pedro J. Leitão, Andreas Rabe, Roland Doerffer, Sebastian Fischer, Timo Stuffler, Rupert Müller, Rudolf Richter, Knut Oppermann, Martin Bachmann, Andreas Müller, Bin Sang, Izabella Walter, Jakub Bieniarz, Andreas Brosinsky, Andreas Eckardt, Uta Heiden, Willem Heldens, Joachim Hill, Patrick Hostert, Hajo Krasemann, Sebastian van der Linden, Wolfram Mauser, Natascha Oppelt, Ruediger Roettgers, Tobias Schneiderhan, Karl Staenz, and Hendrik Wulf. The enmap spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, 7(7):8830–8857, 2015.

Sebastian Hafner, Andrea Nascetti, Hossein Azizpour, and Yifang Ban. Sentinel-1 and sentinel-2 data fusion for urban change detection using a dual stream u-net. *IEEE Geosci. Remote. Sens. Lett.*, 19: 1–5, 2022. doi: 10.1109/LGRS.2021.3119856. URL https://doi.org/10.1109/LGRS.2021.3119856.

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 (CVPR)*, pages 27852–27862, 2024.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

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

Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024.

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.

Jihyeon Lee, Nina R Brooks, Fahim Tajwar, Marshall Burke, Stefano Ermon, David B Lobell, Debashish Biswas, and Stephen P Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17):e2018863118, 2021.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. volume 14, pages 4205–4230, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.

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.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 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*, pages 4088–4099, 2023.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. In *ICML (Position Papers)*, 2024.

David P. Roy, Michael A. Wulder, Thomas R. Loveland, Curtis E. Woodcock, Richard G. Allen, Martha C. Anderson, Dennis Helder, James R. Irons, David M. Johnson, John Kennedy, Tom A. K. Larsen, Jeffrey G. Masek, John R. Schott, Volker C. Radeloff, Crystal L. S. Schaaf, Warren B. Cohen, Christopher J. Crawford, Eileen T. Helmer, Samuel N. Goward, Patrick E. O'Connell, David L. Williams, Robert A. Shuchman, and Ramakrishna R. Nemani. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. Multi3net: segmenting flooded buildings via fusion of multiresolution, multi-sensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019.

Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A satellite time series dataset for crop type identification. In *Proceedings of the International Conference on Machine Learning Time Series Workshop*, volume 3, 2019.

Michael Schmitt, Lloyd H. Hughes, Chunping Qiu, and Xiaoxiang Zhu. 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.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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.

Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, Björn Rommen, Nicolas Floury, Mike Brown, Ignacio Navas Traver, Patrick Deghaye, Berthyl Duesmann, Betlem Rosich, Nuno Miranda, Claudio Bruno, Michelangelo L'Abbate, Renato Croci, Andrea Pietropaolo, Markus Huchler, and Friedhelm Rostan. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. pages 286–303, 2024a.

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*, 2024b.

Yi Wang, Hugo Hernández Hernández, Conrad M. Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 18:321–336, 2025. doi: 10.1109/JSTARS.2024.3493237. URL https://doi.org/10.1109/JSTARS.2024.3493237.

Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Extending global-local view alignment for self-supervised learning with remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2443–2453, 2024.

Western Cape Department of Agriculture and Radiant Earth Foundation. Crop type classification dataset for western cape, south africa. https://doi.org/10.34911/rdnt.j0co8q, 2021. Accessed: [Date Accessed].

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

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.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. 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.

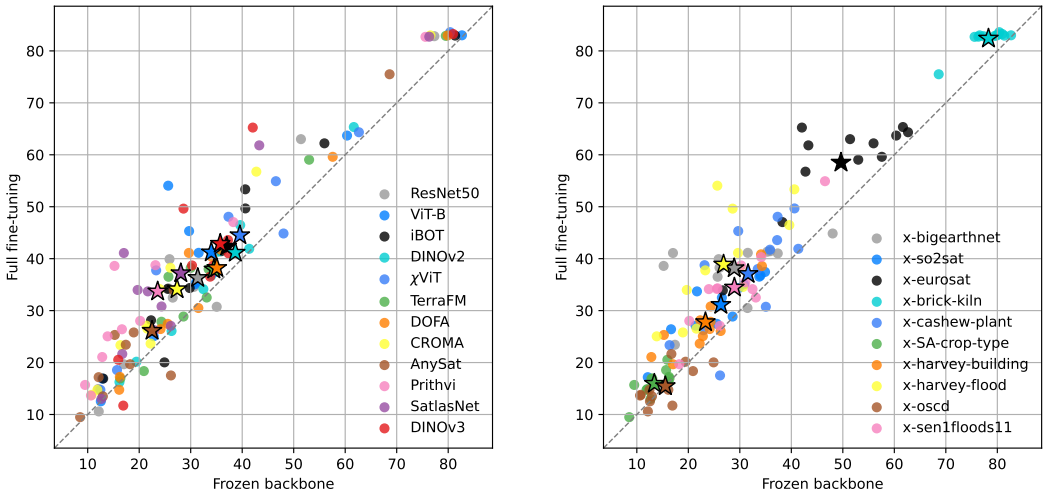
Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, et al. So2sat lcz42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3), 2020.

A ALL RESULTS

For GeoBench datasets, we report the same metrics as in GeoBench. For other datasets on GeoCross-Bench, we adopt the metrics used in previous works. Specifically, for x-sen1floods11, we report mIoU (Bonafilia et al., 2020b; Marsocci et al., 2024; Tseng et al., 2025). For x-OSCD dataset, we report the F1 score (Caye Daudt et al., 2018; Mendieta et al., 2023). For x-harvey-building and x-harvey-flood, we first calculated mIoU and bIoU (Rudner et al., 2019). However, because the classes are highly imbalanced, our initial experiments showed that models can achieve a high mIoU simply by perfectly segmenting the majority class while completely failing on the minority class. To avoid this misleading result, we therefore report only the bIoU for the minority class as our evaluation metric.



(a) Colored by models

(b) Colored by datasets

Figure 4: Performance of the models with frozen backbone (x-axis) vs. full fine-tuning (y-axis) for all pairs of models and datasets. (a) figure shows results colored by models, where stars indicate model's average performance. In figure (b) results are colored by datasets and stars are the average performance on each dataset.

B FROZEN BACKBONES VS. FULL FINE-TUNING

Figure 4 highlights the difference in performance with frozen and non-frozen backbones. For most of the pairs, full fine-tuning is slightly better. In fact, on average, all models are above the $y = x$ line. Only certain dataset-model pairs are below the diagonal (e.g. AnySat on x-cashew-plant).

C PRETRAINING DATASET

The main components are summarized in Table 3. A key part comes from the Satlas Pretrain dataset (Bastani et al., 2023), which includes Sentinel-1 SAR images (using VV and VH polarizations, which are the absolute values of the complex numbers), Sentinel-2 multispectral images (using 10 bands, excluding B01, B09, and B10 which have lower resolution), and NAIP high-resolution RGB aerial photos. The MillionAID dataset (Long et al., 2021) contributes a large volume of RGB aerial images with varied resolutions and image sizes. 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 feature. The Intelinair dataset (Chiu et al., 2020) gives very detailed (0.02m GSD) RGB and Near-Infrared (NIR) aerial images of farms. Using all these different datasets together gives χViT a solid base for pretraining. This helps the model learn features that work well across different kinds of sensors.

Table 3: Overview of datasets used for χViT pretraining.

Dataset	Bands	# Images	GSD	Image Size
Satlas (Bastani et al., 2023) (S1)	Sentinel-1 (VV, VH)	4.5M	20m	512×512
Satlas (Bastani et al., 2023) (S2)	Sentinel-2 (10 bands)	11M	10m	512×512
Satlas (Bastani et al., 2023) (NAIP)	Aerial RGB	5.3M	1m	512×512
MillionAID (Long et al., 2021)	Aerial RGB	2M	0.5-153m	$(170 - 550)^2$
BigEarthNet (Sumbul et al., 2021)	S1 SAR & S2 (10 bands)	0.55M	10m	120×120
Sen12MS (Schmitt et al., 2019)	S1 SAR & S2 (10 bands)	0.18M	10m	256×256
Intelinair (Chiu et al., 2020)	Aerial RGB, NIR	34K	0.02m	320×320

D PRETRAINING DETAILS

The pre-training process is visualized in Figure 2, where we utilized a multi-crop setup with 8 local views (denoted as L_i in the figure) and 2 global views (e.g., G_1, G_2). For the final pretraining of χViT, we processed 400 million samples in total. The AdamW optimizer (Loshchilov and Hutter, 2019) was used with a batch size of 512. Given that the final pretraining did not have a predefined number of iteration steps, as we aimed to train for as long as we could manage, we utilized the Warmup-Stable-Decay (WSD) learning rate scheduler (Hu et al., 2024). This approach allowed for a flexible decay phase, which was initiated for the last 10% of total iterations. The learning rate was linearly warmed up for the initial 30 million samples to a peak of 2.5×10^{-4}, maintained during the stable phase, before the final decay. It's common practice to adjust this peak learning rate in proportion to the batch size (e.g., using the formula $peak\_lr \times batch\_size/256$). The overall loss was a sum of the [CLS] token self-distillation loss and the MIM (masked image modeling) loss, without scaling factors between them.

Design choices. To determine the optimal configuration for χViT, we conducted several experiments for the key design choices. Each experimental configuration was pre-trained for 40 million samples. Model selection was based on the mean Average Precision (mAP) achieved after fine-tuning on 1% of the BigEarthNet dataset. The results of these ablation studies are summarized in Table 4. Based on these experiments, the winning configuration utilized subset sampling for student channels by sampling from teacher channels (employing hierarchical channel sampling as described in (Bao et al., 2024)), shared projection weights for all bands, a shared prediction head for CLS and patch tokens, and a parallel data coefficient of 4 for the BigEarthNet and Sen12MS datasets during pretraining.

Table 4: Ablation study for χViT pretraining design choices. Each configuration was pre-trained for 40M samples. Performance was evaluated by fine-tuning on 1% of BigEarthNet and measuring mAP. PDC refers to the Parallel Data Coefficient.

Subset Sampling	Shared Proj.	Shared Head	PDC (λ)	mAP (%)
✓	✓	✓	**4**	**54.72**
x	✓	✓	4	51.10
✓	x	✓	4	40.44
✓	✓	x	4	46.55
✓	✓	✓	8	51.51

E FINE-TUNING

For models whose input channel count is fixed and smaller than the number of bands in our data, we adapt the first convolutional layer. For Sentinel-2 during training, we average the pretrained first-layer weights across the original input channels to obtain a single-channel kernel, replicate this kernel across all input bands, and divide by the new input channel count. For RGBN evaluation with models pretrained for three-channel input, at inference we modify the first layer by setting the weights of the fourth channel to the mean of the weights of the three original channels.

E.1 HYPERPARAMETERS

We apply a grid search to find the best learning rate and decoder depth. Similar to Tseng et al. (2025) we swept learning rates over the sets $\{1, 3, 6\} \times \{10^{-5}, 10^{-4}, 10^{-3}\}$ for full fine-tuning, and $\{1, 3, 4, 5\} \times \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ for fine-tuning with a frozen encoder. For the UPerNet decoder (Xiao et al., 2018), we scale its width with values from the set $\{1, 2, 3\}$ in our grid. Recognizing that the optimal hyperparameters for a given task were often very similar, if not identical, across different models, we conducted the hyperparameter search on selected models, specifically iBOT and DOFA, across all datasets. We then ranked the configurations and selected the top-ranked hyperparameter set that performed well for both iBOT and DOFA, applying them to the remaining models for that particular task. Tables 5 and 6 show the chosen hyperparameters for each dataset.

Table 5: Chosen hyperparameters for full fine-tuning.

Dataset	Learning rate (LR)	UPerNet width
Sen1Flood11	6×10^{-5}	2
Harvey Segmentation	6×10^{-4}	1
Harvey Change Detection	5×10^{-4}	1
OSCD	3×10^{-4}	2

Table 6: Chosen hyperparameters for fine-tuning with a frozen backbone.

Dataset	Learning rate (LR)	UPerNet width
Sen1Flood11	4×10^{-4}	2
Harvey Segmentation	3×10^{-3}	1
Harvey Change Detection	5×10^{-3}	1
OSCD	1×10^{-3}	2

E.2 FINE-TUNING DETAILS

Across all tasks, we consistently apply the following settings. We use a learning-rate scheduler featuring a 20-epoch linear warmup phase, which is then followed by a cosine decay. This decay period lasts for 30 epochs in classification tasks, and 80 epochs for both segmentation and change detection tasks. We use AdamW optimizer (Loshchilov and Hutter, 2019) for all model trainings. For input normalization, we apply channel-wise mean and standard deviation normalization, clipping the resulting values to the range $[-3, 3]$. Model selection is based on by choosing the checkpoint with the highest validation metric. We set batch sizes to 64 for classification tasks, and to 8 for both segmentation and change detection tasks.

F COMPUTE RESOURCES

Our experiments were performed on three machines: DGX A100 and DGX H100 at Yerevan State University and one HGX H100 node kindly donated by Nebius.ai cloud.

The final pretraining of χViT required 12 days of 8 H100s (96 H100-days). Before the final version we had one more similar run which had a bug in the layer unfreezing code which resulted in a poor performance.

Fine-tuning compute strongly depends both on the model and the dataset. We used 5 seeds for every pair. We also performed hyperparameter search on 4 datasets and 2 models with 27 or 48 combinations of hyperparameters. All these experiments were scheduled with Slurm on A100 and H100 nodes, and we did not track which experiments went to which GPU. In total, we estimate all fine-tuning efforts (including hyperparameter search) used 45 GPU-days.

We also estimate that we wasted another 40 GPU-days on running initial experiments for each baseline model. Many experiments were performed on older versions of the datasets (e.g. original BigEarthNet v1.0, or non-GeoBench versions of datasets) that were excluded from this paper.