Bridging Remote Sensors with Multisensor Geospatial Foundation Models

Boran Han[1] ✉ Shuai Zhang[1] Xingjian Shi[2] * Markus Reichstein[1,3]
[1] Amazon Web Services [2] Boson AI [3] Max-Planck-Institute for Biogeochemistry

Abstract

In the realm of geospatial analysis, the diversity of remote sensors, encompassing both optical and microwave technologies, offers a wealth of distinct observational capabilities. Recognizing this, we present msGFM, a multisensor geospatial foundation model that effectively unifies data from four key sensor modalities. This integration spans an expansive dataset of two million multisensor images. msGFM is uniquely adept at handling both paired and unpaired sensor data. For data originating from identical geolocations, our model employs an innovative cross-sensor pretraining approach in masked image modeling, enabling the synthesis of joint representations from diverse sensors. msGFM, incorporating four remote sensors, upholds strong performance, forming a comprehensive model adaptable to various sensor types. msGFM has demonstrated enhanced proficiency in a range of both single-sensor and multisensor downstream tasks. These include scene classification, segmentation, cloud removal, and pan-sharpening. A key discovery of our research is that representations derived from natural images are not always compatible with the distinct characteristics of geospatial remote sensors, underscoring the limitations of existing representations in this field. Our work can serve as a guide for developing multisensor geospatial pretraining models, paving the way for more advanced geospatial capabilities. Code can be found at https://github.com/boranhan/Geospatial_Foundation_Models

1. Introduction

Geospatial remote sensors exhibit considerable diversity (Figure 1), with reported spatial [44] and feature heterogeneity [58, 64]. Two principal categories emerge based on their imaging mechanisms: optical sensors (e.g., Sentinel-2 [18] and LiDAR) and microwave sensors (e.g., Synthetic-aperture radar [21]). These sensors vary significantly in their observation methods and capabilities. Optical remote sensing captures reflected and absorbed electromagnetic radiation in the visible and near-infrared spectrum, yielding high-resolution imagery and surface property information. Conversely, microwave remote sensing operates at longer wavelengths, penetrating clouds and vegetation to reveal subsurface features and structural properties [41] (Figure 1).

A multisensor fusion approach combines the strengths of both optical and microwave remote sensing, offering a more comprehensive and accurate understanding of the Earth's surface [51]. By integrating data from multiple sensors, researchers can leverage the complementary nature of optical and microwave data to overcome limitations and obtain a more complete picture. For instance, combining optical and microwave data can help estimate soil moisture content, which is crucial for ecosystem management [24, 31]. Multisensor fusion also enhances the accuracy of topographic mapping by incorporating both surface features captured by optical sensors and elevation information derived from microwave sensors. Numerous multisensor fusion deep learning models have been proposed for individual tasks, such as cloud removal [28, 65], biomass estimation [23] and land-cover segmentation [7, 29]. These studies substantiate the enhancement in performance achievable by geospatial models incorporating multisensor modalities.

Despite these important synergies, most geospatial pretrained models focus predominantly on a single modality [15, 38, 39, 57, 60]. While studies like Liu et al. [34], Chen and Bruzzone [9] and Scheibenreif et al. [50] employ Sentinel-2 and SAR for pretraining via contrastive learning, these methodologies are inherently limited by the need to paired sensor modalities. This limitation restricts the efficient utilization of the abundant unpaired sensor modalities that are prevalently available in real-world scenarios. By establishing a multisensor pretrained model scalable to both paired and unpaired sensors, a unified framework for analyzing multisensor remote sensing data can be provided. Such a model can be fine-tuned or used as a feature extractor to interpret multisensor data effectively.

Therefore, our paper develops a novel multisensor geospatial pretraining model that can leverage many sensor modalities, paired or not. Additionally, our paper seeks to address several unexplored questions in the realm of multisensor geospatial models. A natural inquiry arises: *How*

*Work done while at Amazon Web Services
✉ Corresponding author

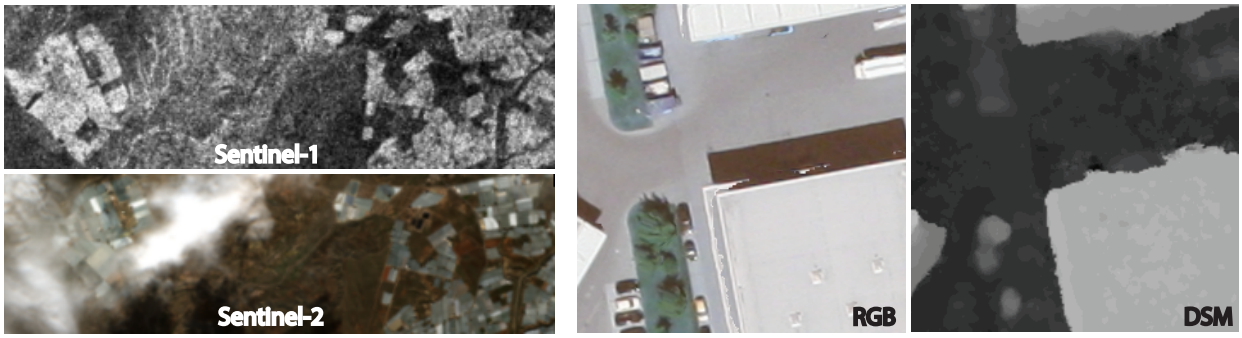

Figure 1. Examples of four sensor modalities: SAR, Sentinel 2, RGB, DSM. Here, each pair of {SAR & Sentinel-2} and {RGB & DSM} are colocated on the same geolocation respectively. In the example of Sentinel-2, only blue, green, red bands are shown for the convenience of visualization.

can joint representations between corresponding sensors be learned by employing masked image modeling techniques? In geospatial tasks within the RGB domain, it is typical to leverage pretrained backbones on ImageNet [48, 61] or to utilize features distilled from such models [39]. Given this, we inquire, *Does leveraging or distilling features from established vision models enhance multisensor geospatial pretraining?* Lastly, a practical concern emerges: *How can multisensor heterogeneity be mitigated during pretraining?* Addressing these challenges is crucial for developing geospatial pretrained models capable of handling diverse sensor data. Our contributions can be summarized as follows:

- We introduce a novel cross-sensor paradigm, msGFM, for joint representation learning. This paradigm harmonizes diverse representations and empowers multisensor models to effectively discern the complex relationships between corresponding sensors.
- We introduce a high-performing pretrained model, cultivated from a comprehensive multisensor pretraining dataset encompassing over 2 million images. This model adeptly amalgamates four sensor modalities: RGB images, Sentinel-2, SAR, and DSM, demonstrating superior performance across several important downstream tasks.
- We demonstrate the synergistic advantages of incorporating multiple sensor modalities in pretraining, as opposed to focusing on single-sensor approaches. In addition, our work includes a thorough analysis of the model, offering practical insights and strategies for achieving optimal performance in multisensor foundation models.

2. Related Work

Geospatial pretraining. Geospatial technologies are becoming increasingly essential for applications, such as planning, monitoring and disaster response [22, 26, 46]. As pretrained models continue to revolutionize the fields of vision and language, their potential in the geospatial sphere is becoming increasingly evident. These models have demonstrated remarkable prowess in enhancing the efficacy of deep learning models when applied to downstream tasks [4, 15, 38, 39, 42]. The geospatial domain has seen the emergence of two main approaches for self-supervised pretraining paradigms. The first centers around the use of contrastive learning [4, 34, 38]. In this technique, the loss function is crafted to incentivize the model to draw similar or positive pairs closer together in the embedding space while pushing dissimilar or negative pairs further apart [8]. However, identifying appropriate augmentations for contrastive methods presents a significant challenge. Certain augmentations in geospatial images, which significantly alter the image's intensity, can lead to undesirable outcomes [42]. Various implementations of pretraining with contrastive learning incorporate temporal and spectral augmentation [38], while others apply a colorization objective [59]. Although works such as Liu et al. [34], Chen and Bruzzone [9] and Scheibenreif et al. [50] treat colocalized Sentinel-2 and SAR as positive pairs, these approaches are restricted to these two or more pairing sensor modalities and doesn't efficiently leverage the wide range of unpaired sensor modalities. Given these augmentation constraints [42], alternative methods have been developed, employing Masked Image Modeling (MIM) [15, 39, 57], relying on simple spatial augmentations such as flipping and cropping. MIM not only requires less stringent augmentations but also outperforms its contrastive learning counterparts [15, 39, 57]. However, most prior studies focus on remote sensing imagery in the visible spectrum or employ a single sensor modality [15, 38, 39, 60]. Alternatively, they are confined to *two or more paired* sensors due to the inherent limitations of contrastive learning [9, 34, 50]. In this work, we develop our pretraining objective based on the masked image modeling approach, akin to [27, 63]. We demonstrate that our model can be pretrained with four sensor modalities, taking advantage of the unpaired sensor.

Multi-source learning in language and vision communities. Multi-source learning is a prevalent strategy

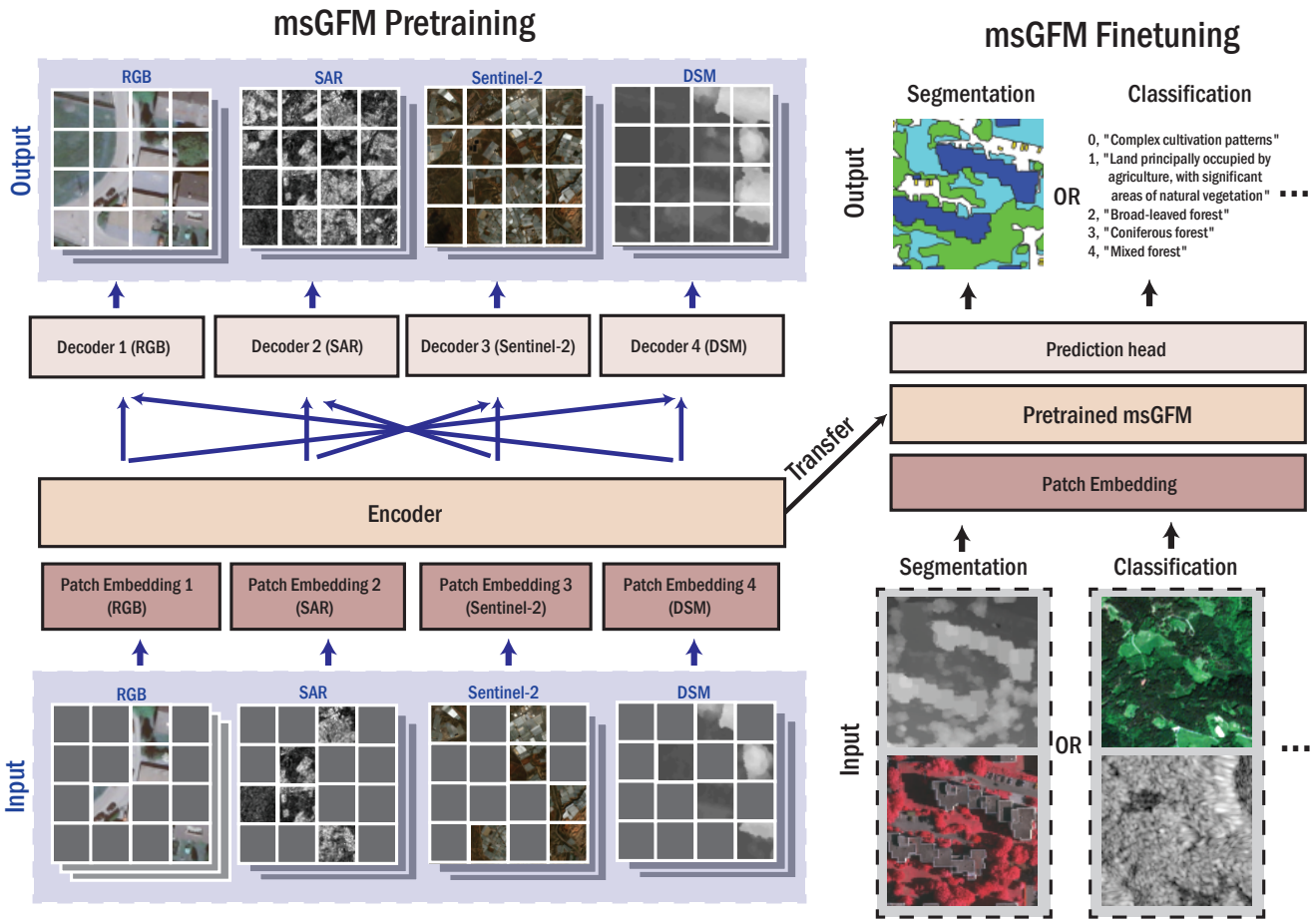

Figure 2. Overview diagram of msGFM. Each sensor is fed through a separate patch embedding layer (Section 3.1) and through the same encoder. For reconstruction, separate decoders are used. If the sensors are paired, there's a chance that our model will reconstruct the corresponding paired sensor instead of itself (Section 3.2). Other best practices can be found in Section 3.4. In the finetuning stage, the pretrained encoder (msGFM) is transferred to different downstream applications with different prediction heads. In Appendix C.2, we discuss the usage of patch embedding in the downstream finetuning.

when handling multi-modal [54, 66] and multitask challenges [10, 11] in both the language and vision domains [2, 3, 5, 12, 16, 32, 33]. This technique exploits data from diverse sources to bolster the learning process and enhance model performance. A notable example is multilingual pretrained models, such as XLM [32] and its derivatives [13, 16]. These models utilize multilingual datasets, pretraining them on a large scale to generate unsupervised cross-lingual representations [16, 32]. This approach enables the models to develop a unified representation across multiple languages, thereby enhancing their performance on cross-lingual tasks. Furthermore, the batching strategy has been identified as an essential aspect of creating generalizable representations and preventing collapse in multilingual models [1, 2]. Simultaneously, the Mixture-of-Experts (MoE) strategy [53] has been utilized to enhance multi-source learning in both multitask learning [3, 12, 33] and language-vision pretraining [54, 66].

In the specific context of multisensor geospatial pretraining, heterogeneity can originate from the use of different sensor types (e.g., optical, microwave) or different platforms (e.g., various satellites). Properly addressing this heterogeneity is crucial as it can significantly influence the performance of the pretraining model [2]. To meet this challenge, we draw inspiration from works in vision [47], language [2, 32] and vision-language models [66]. We incorporate techniques such as cross-sensor representation learning into our msGFM.

3. Cross-sensor geospatial pretaining

In this section, we present the multisensor pretraining paradigm. Following [39], we employ SIMMIM [63] using a Swin Transformer [35, 36] as a backbone. Figure 2 presents an overview of the cross-sensor geospatial pretrain-

ing methodology.

3.1. Input representation

Distinct embedding layers for each sensor. We consider N total types of sensor modalities, with each sensor having a corresponding number of channels, denoted as $\{C_i\}_{i=1...N}$. Taking into account the unique number of channels associated with each sensor (some examples shown in Table 1), we utilize individual patch embeddings tailored to each specific sensor. This approach allows the model to efficiently process and learn from the distinct characteristics of various sensor modalities.

The patch embedding is primarily obtained from convolution layers. To elaborate, for the image from the i-th sensor, $\boldsymbol{I} \in \mathbb{R}^{W\times H\times C_i}$. Here, W and H represent the width and height of the images. We first apply the convolution layers, $\{f_i\}_{i=1...N} : \mathbb{R}^{W\times H\times C_i} \rightarrow \mathbb{R}^{W\times H\times C_e}$. The function f_i represents the convolution neural network layers for images from the i-th sensor. In our work, C_e is the same for all sensors. Subsequently, the output is segmented into square patches of size P, yielding $\boldsymbol{T}_i \in \mathbb{R}^{L\times P^2 C_e}$, where L represents the total number of patches. To effectively manage the channel heterogeneity inherent in different sensor modalities, each sensor modality is processed through its own trainable embedding layer. This step standardizes the representation dimensions before they are input into the shared encoder.

3.2. Cross-sensor pretraining

Shared encoder for all sensor modalities. The patches obtained from $\{f_i\}_{i=1...N}$, $\boldsymbol{T}_i \in \mathbb{R}^{L\times P^2 C_e}$, will then be masked and fed through the encoder. The masking strategy employed in our approach is the same as those used in [63]. By having separate patch embedding layers (f_i) for each sensor, the model can learn the unique characteristics of each sensor modality. The learned embeddings from all sensors are then integrated through the same encoder, enabling the model to effectively learn joint representations and handle multisensor geospatial data.

Separate decoder for each sensor and cross sensor prediction. Collecting data from different sensors for the same geo-location is a common practice in the geospatial domain. Learning joint representations of such multisensor data can prove beneficial for various downstream tasks. Although contrastive learning has demonstrated promise in learning effective representations, its performance may be limited due to the lack of suitable data augmentations for remote sensing images [42]. To address this issue, we propose employing cross-sensor strategies in the context of MIM to learn joint representations for multisensor geospatial data. For instance, when the model is fed with masked images from DSM, it can predict the masked patches of itself or the corresponding images from RGB. An example pair of DSM and RGB images is shown in Figure 1 in the two panels on the right. This encourages the model to align the different sensor representations. Accordingly, our model incorporates different decoders for each sensor.

Specifically, if there exists a pair of images from the i-th sensor and j-th sensor, $\{\boldsymbol{I}_i \in \mathbb{R}^{W\times H\times C_i}, \boldsymbol{I}_j \in \mathbb{R}^{W\times H\times C_j}\}$, the model processes the masked image as follows:

$$\begin{aligned} \boldsymbol{I}'_i = D_i(En(f_i(\boldsymbol{I}_i))) \text{ , or } \boldsymbol{I}'_j = D_j(En(f_i(\boldsymbol{I}_i))) \\ \text{and } \boldsymbol{I}'_j = D_j(En(f_j(\boldsymbol{I}_j))) \text{ , or } \boldsymbol{I}'_i = D_i(En(f_j(\boldsymbol{I}_j))) \end{aligned} \quad (1)$$

where $En : \mathbb{R}^{L\times P^2 C_e} \rightarrow \mathbb{R}^{L\times C_m}$ is the shared encoder, and C_m is the embedding dimension of the final layer in the encoder. I'_i and I'_j are the predicted i-th and j-th sensor type respectively. $D_i : \mathbb{R}^{L\times C_m} \rightarrow \mathbb{R}^{W\times H\times C_i}$ is the decoder to reconstruct the i-th sensor type. Equation 1 shows that the predicted output of the pretraining model will either reconstruct itself or its paired sensor images.

If there's no paired sensor in the pretraining dataset, it will construct itself in the conventional way:

$$\boldsymbol{I}'_i = D_i(En(f_i(\boldsymbol{I}_i))) \quad (2)$$

This approach capitalizes on the inherent relationship between different sensors observing the same location, enabling the model to capture complementary information. Furthermore, it provides flexibility in handling scenarios where no paired sensors are available, allowing for enhanced adaptability in choosing the pretraining dataset. This is particularly advantageous given that multisensor geospatial datasets are less prevalent than single-sensor datasets.

3.3. Pretraining data.

Our multisensor pretraining data, GeoPile-2, is composed of four sensor modalities, amassed to a total of 2 million images through the inclusion of additional geospatial data. The detailed composition of GeoPile-2 is presented in Table 1. Specifically, GeoPile-2 incorporates SEN12MS [52], a dataset enriched with paired SAR and Sentinel-2 satellite images from all meteorological seasons, to augment data diversity. All sensors in this dataset are ortho-rectified [52]. Additionally, we have integrated DSM and RGB images from the MDAS dataset [29], resized to a dimension of 384.

It is important to note that while the Sentinel-2 modality includes RGB channels as part of its imaging band, we have distinguished and separated this RGB modality due to its extensive dataset that surpasses the scope of Sentinel-2. This dataset exhibits a wide range of Ground Sample Distances (GSD) and high feature entropy. These attributes, beyond just the imaging band, have been proven to be influential in pretraining, as evidenced in studies like [15, 39]. Our observations indicate that excluding the RGB modality from our pretraining dataset (GeoPile-2) leads to a decrease in

Dataset	# Images	GSD	Sensor modality	# Channels	paired sensors?
GeoPile [39]	600K	0.1m - 30m	RGB[a]	3	✗
MillionAID [37]	1M	0.5 - 153m	RGB[a]	3	✗
SEN12MS [52]	320K	10m	SAR / sentinel-2	2/14	✓
MDAS [29]	40K	0.1m - 10m	DSM/ RGB[b]	1/3	✓

Table 1. Breakdown of datasets of our pretraining data. We gather approximately 2M samples from a combination of labeled and unlabeled satellite imagery with various ground sample distances and sensor modalities. GeoPile and MillionAID are not sourced from a single sensor; instead, they amalgamate sensor images from an array of satellites, including NAIP, GeoEye, WorldView, QuickBird, IKONOS, and SPOT satellites, among others. MDAS is derived from airborne sources [29]. For more in-depth details regarding the RGB, please refer to Appendix A.2

effectiveness, as opposed to when it is included (see Appendix A.2). An enhanced version of GeoPile-2, which includes RGB data from sources like GeoPile [39] and MillionAID [37], shows improved performance across the seven downstream tasks specified in GFM [39], under similar experimental conditions (detailed in Appendix A.1). Further optimization of GeoPile-2-RGB was achieved through experiments with various datasets (refer to Appendix A.1).

3.4. Best Practices in Pretraining

A shared encoder can present challenges when it comes to efficiently learning each sensor's representation. To tackle this issue, we propose integrating the sparsely gated Mixture of Experts (MoE) approach [53] to replace MLP layers within the encoder. Our pretraining loss function, L, combines L1 loss [27, 63] for reconstruction (i.e., MIM loss) and auxiliary losses [30, 47]: $L = L_{\text{MIM}} + \lambda L_{\text{auxiliary}}$, where λ represents the weight for auxiliary losses. In practice, we use $\lambda = 0.01$.

We sequentially load all sensor data in our model. This approach ensures that our model's optimization spans all tasks. Specifically, each batch in our model is constituted as a set: $\boldsymbol{I} \in \mathbb{R}^{W \times H \times C_i}{}_{i=1...N}$. Such a methodology is commonly utilized in multitask learning, aiming to forge a unified representation across diverse tasks during the training process, despite their distinct learning objectives. Our investigations reveal that this strategy is equally effective in the context of multisensor geospatial pretraining. Furthermore, considering the unique imaging mechanisms of these sensors, we choose to initiate pretraining *from scratch*, as detailed in Section 4.4.

4. Experiments

Experimental Settings. All of our experiments are conducted using a Swin-base architecture [36] with a patch size of 16×16 pixels and 8 experts. The models are pretrained for either 100 epochs for ablation studies or 800 epochs to achieve optimal results and maintain comparability with state-of-the-art methods. When specified, 1% BigEarthNet (BEN) [56] and 1% SEN12MS-CR are also employed for ablation studies. We utilize 8 NVIDIA V100 GPUs with a batch size of 2048 (128 per GPU) and an image size of 192×192. All pretraining settings follow the configurations described in [39]. Detailed pretraining settings and pretraining cross-sensor reconstruction visualization can be found at Appendix B.1 and Section 4.1 respectively.

Downstream Evaluation. Upon completion of the pretraining, we fine-tune and assess the model on a diverse range of downstream multisensor datasets. This aims to provide a comprehensive understanding of the model's performance potential across various tasks. Table 2 provides an overview of the downstream evaluation tasks, together with their respective sensor modalities. Among these tasks, the use of multisensor data can enhance the performance of land classification and segmentation. Meanwhile, cloud removal is inherently dependent on multisensor modalities and cannot be effectively tackled without them. Although pansharpening requires one optical sensor, it relies heavily on multi-spectral images.

4.1. Visualizing reconstruction quality

To demonstrate this methodology, we provide several examples in Figure 3. These instances visually illustrate our cross-sensor pretraining approach, highlighting the ability of RGB to self-reconstruct, as well as the excellent cross reconstruction capabilities between DSM and RGB images. However, self-reconstruction and cross reconstruction involving SAR images pose some challenges, as we use unprocessed, noisy SAR images. Due to the structure of MIM, which involves an encoder and a lightweight reconstruction decoder, only low-frequency components of the images are reconstructed [27, 63], making the SAR reconstruction slightly difficult. Specifically, we visualized the statistics [6] and SSI [55] value before and after the reconstruction for both HV and VV bands (Figure 5).

4.2. Geospatial downstream evaluation

4.2.1 Scene classification.

One prevalent remote sensing application is classification. We evaluate our pretraining model on BEN, a dataset extensively used in other literature [8, 15, 38, 39, 62]. BEN [56]

Dataset	# Application	GSD	Sensor modality	# Channels
Big Earth Net [56]	Scenes classification	10m - 60m	SAR / Sentinel-2	2/14
Vaihingen [49]	Land segmentation	0.09m	DSM / RGB	1/3
SEN12MS-CR [19]	cloud removal	10 - 60m	SAR / Sentinel-2	2/14
SpaceNet	Pan-sharpening	0.1m - 10m	WorldView 3	8

Table 2. Downstream tasks. It covers various use cases in geospatial domain, with a range of ground sample distances and sensor modalities.

Methods	10% BEN	100% BEN	SEN12MS-CR			SpaceNet		Vaihingen
	mAP (↑)	mAP (↑)	MAE (↓)	SAM (↓)	SSIM (↑)	PSNR (↑)	SSIM (↑)	mIOU (↑)
SeCo [38]	82.6	87.8	-	-	-	-	-	68.9
SatMAE [15]	82.1	-	-	-	-	22.742	0.621	70.6
MoCoV2 [8]	-	89.3	-	-	-	-	-	-
DINO-MC [62]	84.2	88.6	-	-	-	-	-	-
GFM [39]	86.3	-	-	-	-	22.599	0.638	75.2
Random	82.6	86.2	0.048	14.78	0.572	21.825	0.594	67.0
IN-22k [36]	85.7	89.5	-	-	-	21.655	0.612	74.7
msGFM	**87.5**	**92.9**	**0.026**	**4.87**	**0.842**	**22.850**	**0.668**	**75.8**

Table 3. Quantitative results of all the downstream tasks (Table 2) from msGFM (ours) compared to other pretrained models. Results are replicated from the previous reports. Random: random initialization.

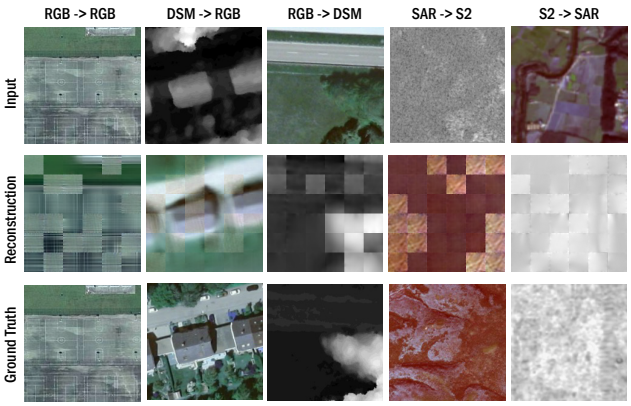

Figure 3. Examples of cross-sensor pretraining. The first row represents the input before masking, the second row depicts the reconstruction, and the third row shows the ground truth.

is a large-scale imbalanced remote sensing dataset specifically designed for multi-label classification tasks. The data includes pairs of 12-band Sentinel-2 images along with their corresponding 2-band SAR images. We employ the data split and 19-class evaluation, as is standard in the literature [15, 38, 39, 42]. In Table 3, we report the mean average precision (mAP) results on BEN for all methods. Our model can provide robust performance against other pretraining methods [8, 15, 36, 38, 39, 62], including ImageNet-22k [36]. More results of the random initialization and ImageNet initialization can found in previous studies [15, 39].

Methods	MAE (↓)	SAM (↓)	SSIM (↑)
SAR-Opt-cGAN [25]	0.043	15.49	0.764
DSen2-CR [40]	0.031	9.47	0.874
GLF-CR [65]	0.027	7.65	**0.885**
msGFM	**0.026**	**4.87**	0.842

Table 4. Quantitative results of cloud removal, compared to existing models that are specially designed for cloud removal. Results are replicated from the original paper.

4.2.2 Cloud removal

The majority of optical observations acquired via spaceborne Earth imagery are affected by clouds, presenting challenges in reconstructing cloud-covered information in current studies. While optical imagery is impacted by adverse weather conditions and the lack of daylight, SAR sensors remain unaffected, offering a valuable source of complementary information. Consequently, performing cloud removal tasks without SAR data significantly degrades task performance [65]. We evaluate our model on SEN12MS-CR [19]. The results, presented in Table 4, show promising performance in Spectral Angle Mapper (SAM) and Mean Absolute Error (MAE), outperforming existing cloud removal models [25, 40, 65]. Additionally, the Structural Similarity Index Measure (SSIM) metric yields comparable results to these methods. Our multisensor pretraining approach, incorporating SAR data, facilitates effective cloud removal. In contrast, other geospatial pretraining models that rely solely on op-

tical data fall short in demonstrating their cloud removal capabilities, as shown in Table 3.

4.2.3 Pan-sharpening.

Pansharpening, akin to super-resolution, involves combining a high-resolution grayscale panchromatic image with the color information from a low-resolution multispectral image to generate a high-resolution color image. For this assessment, we utilized the SpaceNet2 dataset, following the methods in [39]. We juxtaposed the performance of our model with a series of baseline models, measuring the outcomes using peak signal-to-noise ratio (PSNR) and SSIM. As illustrated in Table 3, our model demonstrates superior performance over its competitors. Notably, the SpaceNet dataset, which comprises images from the WorldView-3 satellite—a source not included in our pretraining data—demonstrates the strong transferability of our pretrained model across diverse sensors.

4.2.4 Segmentation

Segmentation is another popular remote sensing application for enabling automated extraction of building footprints or land cover mappings over wide regions. We therefore conduct experiments on this task using Vaihingen [49], which is an urban semantic segmentation dataset collected over Vaihingen, Germany at a GSD of 0.9m. The experiment settings are the same as [39]. We report the intersect of union (IoU) segmentation results for all methods in Table 3. Our approach is able to provide the best result.

4.3. Comparison with single sensor pretraining.

To underscore the pivotal role of multiple sensor modalities in pretraining and validate that our method leads to multisensor synergy, we compare our multisensor pretraining approach using GeoPile-2 with models pretrained on only one sensor modality (i.e., either SAR or Sentinel-2) from SEN12MS [52]. We assess the performance of these models on the BEN [56] and SEN12MS-CR [19], employing both sensors individually and in combination. Figure 4 highlights two advantages of our model: (1) The multisensor pretraining model consistently outperforms models pretrained with a single sensor modality, as indicated by superior performance across all columns when the sensor modality is fixed. (2) Using both sensors for tasks like land use classification and cloud removal leads to enhanced performance, demonstrated by higher accuracy across all rows when the pretraining data is fixed.

The second advantage can be attributed to the complementary data provided by both sensors. Sentinel-2 images are widely used for land use and land cover classification tasks due to their rich spectral information. In contrast, SAR images offer critical insights for identifying water bodies and urban structures, capturing the radar backscatter properties of the Earth's surface. We observe that different tasks have various responses to multisensor data, a phenomenon also evident in the cloud removal task previously studied [65].

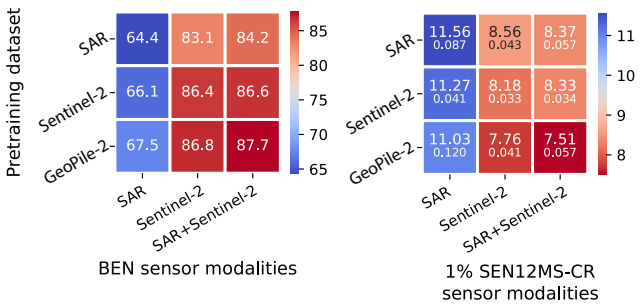

Figure 4. Comparison of our multisensor approach with single modality pretraining on 10% of the BEN dataset (left) using mAP (↑) and 1% of SEN12MS-CR (right) using SAM (↓). Given the reduced dataset size for cloud removal (1% of SEN12MS-CR), we conduct the experiment in three replicates and report both the mean (top line in each cell) and standard deviations (bottom line in each cell).

Notably, even when evaluating the BEN dataset using only the Sentinel-2 modality, our method still achieves superior results compared to other pretraining methods (86.8%). This demonstrates that the improvement of msGFM is not solely due to an increase in sensor modality within the downstream tasks.

4.4. Pretraining from scratch performs better.

In geospatial tasks, especially within the RGB spectrum, it is common to use backbones pretrained on ImageNet [48, 61], or to leverage features distilled from such models [39]. Consequently, we assess the efficacy of leveraging established vision pretrained models for multisensor geospatial pretraining. For fair comparisons, all experiments are trained for 100 epochs.

In one experiment, we extract intermediate features following the methodology in [39] and benchmark them against embeddings obtained from ImageNet-22k. We also conduct a parallel experiment utilizing the CLIP model [45], known for its robust multimodal representation learning.

The results, shown in Table 5, indicate that feature distillation from ImageNet-22k outperforms that from CLIP [45] in terms of performance. Additionally, we explore the EVA method [20], which deviates from traditional MIM approaches like MAE [27] by reconstructing CLIP features of masked patches rather than the patches themselves. Contrary to expectations, EVA, despite its established superiority in other contexts, does not perform as well in our downstream evaluations. This indicates that CLIP features [45] may face a significant domain gap when applied to multisensor

Methods	1% BEN	10% BEN	SEN12MS-CR			SpaceNet		Vaihingen
	mAP (↑)	mAP (↑)	MAE (↓)	SAM (↓)	SSIM (↑)	PSNR (↑)	SSIM (↑)	mIOU (↑)
Distilled from ImageNet22K [17]	79.4	86.4	0.035	6.42	0.726	22.107	0.621	72.9
Distilled from CLIP [45]	76.6	83.8	0.051	8.96	0.707	22.559	0.674	69.3
Reconstruct CLIP (EVA [20])	73.5	80.6	0.053	9.96	0.689	21.778	0.591	65.7
From scratch	**80.9**	**87.2**	**0.026**	**5.04**	**0.821**	**22.742**	**0.677**	**74.8**

Table 5. Distillation from other pretraining model vs pretraining from scatch

Pretraining strategies		Cross sensor	1% BEN	10% BEN	SEN12MS-CR			SpaceNet		Vaihingen
MoE	cross-sensor	percentage	mAP (↑)	mAP (↑)	MAE (↓)	SAM (↓)	SSIM (↑)	PSNR (↑)	SSIM (↑)	mIOU (↑)
✗	✗	0%	78.3	86.2	0.038	8.19	0.735	22.333	0.589	72.8
✓	✗	0%	78.5	86.2	**0.026**	5.11	0.767	22.528	0.637	73.4
✗	✓	50%	**80.7**	86.9	0.036	8.67	0.753	22.518	0.611	**74.6**
✓	✓	100%	80.5	86.8	**0.026**	**4.96**	0.789	22.634	0.649	74.4
✓	✓	50%	**80.9**	**87.5**	**0.026**	5.04	**0.821**	**22.742**	**0.677**	**74.8**

Table 6. Quantitative results of msGFM, with and without MoE/cross sensor reconstruction.

geospatial data.

Conversely, the highest accuracy for multisensor geospatial pretraining is achieved when models are trained from scratch. The lower performance of distillation methods is attributed to the pronounced domain gap between natural images and geospatial-specific sensors. Furthermore, distillation inherently limits the student model's performance to that of the teacher model [39]. This domain gap arises from fundamental differences in the physical mechanisms of optical and microwave remote sensing: while optical remote sensing relies on the reflection and absorption of electromagnetic radiation, microwave remote sensing operates based on the scattering, penetration, and dipole-interference of microwaves [21]. Given that natural images are primarily captured by optical sensors, this leads to a considerable domain discrepancy.

This finding highlights the need for robust foundation models tailored to the geospatial domain, capable of accommodating diverse sensor data and improving multisensor task performance.

4.5. Ablation studies

In the proposed msGFM model, we incorporate both cross-sensor pretraining paradigms and the Mixture of Experts (MoE). In an ablation study, we present the results when either MoE or cross-sensor pretraining is omitted. As shown in Table 6, removing MoE from the model results in similar performance on the some datasets, while other tasks see a more substantial decrease. This uneven response across different tasks aligns with observations made in several previous multi-modal studies [66]. On the other hand, removing the cross-sensor paradigm leads to a consistent performance decline across all tasks.

It is natural to question how the proportion of sensor crossing affects performance. To explore this, we perform an ablation study on the percentage of sensors subject to cross-reconstruction. Our results suggest that a sensor crossing rate of 50% provides slightly superior outcomes compared to a rate of 100%. This indicates that the optimal sensor crossing strategy maintains a balance between the benefits of cross-reconstruction and the retention of sensor-specific information, consequently enhancing performance across a diverse range of geospatial tasks.

5. Conclusion

We introduce a multisensor pretraining model that leverages a novel cross-sensor paradigm to facilitate joint representation learning. This approach adeptly captures the intricate relationships between corresponding sensors. Built on a comprehensive multisensor dataset of over 2 million images, our model showcases outstanding performance across a variety of multisensor downstream tasks.

Looking ahead, there's potential to augment the model's utility for downstream tasks where temporal information plays a pivotal role, such as in predicting ecosystem changes. The integration of temporal data holds immense value but introduces considerable challenges, primarily due to the significant increase in pre-training costs associated with incorporating temporal elements. Thus, the effective amalgamation of spatial and temporal information into a pretrained model demands more than just the inclusion of data; it necessitates a profound rethinking of methodological design.

References

[1] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse, 2020. 3

[2] Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[3] Raquel Aoki, Frederick Tung, and Gabriel L. Oliveira. Heterogeneous multi-task learning with expert diversity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2022. 3

[4] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David B. Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *CoRR*, abs/2011.09980, 2020. 2

[5] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders, 2022. 3, 12

[6] Bernhard Bauer-Marschallinger, Senmao Cao, Claudio Navacchi, Vahid Freeman, Felix Reuß, Dirk Geudtner, Björn Rommen, Francisco Ceba, Paul Snoeij, Evert Attema, Christoph Reimer, and Wolfgang Wagner. The normalised sentinel-1 global backscatter model, mapping earth's land surface with c-band microwaves. *Scientific Data*, 8, 2021. 5

[7] Keumgang Cha, Junghoon Seo, and Yeji Choi. Contrastive multiview coding with electro-optics for SAR semantic segmentation. *CoRR*, abs/2109.00120, 2021. 1

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 2, 5, 6, 12

[9] Yuxing Chen and Lorenzo Bruzzone. Self-supervised sar-optical data fusion of sentinel-1/-2 images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1, 2

[10] Yixin Chen, Shuai Zhang, Boran Han, and Jiaya Jia. Lightweight in-context tuning for multimodal unified models, 2023. 3

[11] Yixin Chen, Shuai Zhang, Boran Han, Tong He, and Bo Li. Camml: Context-aware multimodal learner for large models, 2024. 3

[12] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik Learned-Miller, and Chuang Gan. Mod-squad: Designing mixture of experts as modular multi-task learners, 2022. 3

[13] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland, 2022. Association for Computational Linguistics. 3

[14] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world, 2018. 12, 13

[15] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *arXiv preprint arXiv:2207.08051*, 2022. 1, 2, 4, 5, 6, 12

[16] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. 3

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8, 12, 13

[18] 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, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120: 25–36, 2012. The Sentinel Missions - New Opportunities for Science. 1

[19] Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 6, 7

[20] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 7, 8

[21] Gianfranco Fornaro and Vito Pascazio. Chapter 20 - sar interferometry and tomography: Theory and applications. In *Academic Press Library in Signal Processing: Volume 2*, pages 1043–1117. Elsevier, 2014. 1, 8

[22] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang (Bernie) Wang. Prediff: Precipitation nowcasting with latent diffusion models. In *Advances in Neural Information Processing Systems*, pages 78621–78656. Curran Associates, Inc., 2023. 2

[23] Sujit Madhab Ghosh and Mukunda Dev Behera. Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest. *Applied Geography*, 96:29–40, 2018. 1

[24] Julia Gottfriedsen, Max Berrendorf, Pierre Gentine, Markus Reichstein, Katja Weigel, Birgit Hassler, and Veronika Eyring. On the generalization of agricultural drought classification from climate data. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. 1

[25] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729, 2018. 6

[26] Boran Han and Jeremy Vila. A robust end-to-end method for parametric curve tracing via soft cosine-similarity-based objective function. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision (ICCV) Workshops, pages 2453–2463, 2021. 2

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 2, 5, 7

[28] Wei He, Danfeng Hong, Giuseppe Scarpa, Tatsumi Uezato, and Naoto Yokoya. *Multisource Remote Sensing Image Fusion*, chapter 10, pages 136–149. John Wiley Sons, Ltd, 2021. 1

[29] J. Hu, R. Liu, D. Hong, A. Camero, J. Yao, M. Schneider, F. Kurz, K. Segl, and X. X. Zhu. Mdas: a new multimodal benchmark dataset for remote sensing. *Earth System Science Data*, 15(1):113–131, 2023. 1, 4, 5, 14

[30] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, Joe Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. Tutel: Adaptive mixture-of-experts at scale. *CoRR*, abs/2206.03382, 2022. 5

[31] Martin Jung, Markus Reichstein, Christopher Schwalm, Chris Huntingford, Stephen Sitch, Anders Ahlström, Almut Arneth, Gustau Camps-Valls, Philippe Ciais, Pierre Friedlingstein, Fabian Gans, Kazuhito Ichii, Atul Jain, Etsushi Kato, Dario Papale, Ben Poulter, Botond Raduly, Christian Rödenbeck, Gianluca Tramontana, and Ning Zeng. Compensatory water effects link yearly global land co2 sink changes to temperature. *Nature*, 541, 2017. 1

[32] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019. 3

[33] Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design, 2022. 3

[34] Chenfang Liu, Hao Sun, Yanjie Xu, and Gangyao Kuang. Multi-source remote sensing pretraining based on contrastive self-supervised learning. *Remote Sensing*, 14(18), 2022. 1, 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 3, 12

[36] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 6

[37] 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. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021. 5, 12

[38] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *CoRR*, abs/2103.16607, 2021. 1, 2, 5, 6, 12

[39] Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Gfm: Building geospatial foundation models via continual pretraining, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 12, 13

[40] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020. 6

[41] Z. N. Musa, I. Popescu, and A. Mynett. A review of applications of satellite sar, optical, altimetry and dem data for surface water modelling, mapping and parameter estimation. *Hydrology and Earth System Sciences*, 19(9):3755–3769, 2015. 1

[42] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *CoRR*, abs/1911.06721, 2019. 2, 4, 6

[43] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip, 2023. 12

[44] Bingwen Qiu, Canying Zeng, Chongcheng Cheng, Zhenghong Tang, Jianyang Gao, and Yinpo Sui. Characterizing landscape spatial heterogeneity in multisensor images with variogram models. *Chinese Geographical Science*, 24: 317–327, 2013. 1

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 7, 8

[46] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Mr Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195, 2019. 2

[47] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts, 2021. 3, 5

[48] Vladimir Risojevic and Vladan Stojnic. The role of pretraining in high-resolution remote sensing scene classification. *CoRR*, abs/2111.03690, 2021. 2, 7

[49] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012. 6, 7

[50] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1421–1430, 2022. 1, 2

[51] Michael Schmitt, Florence Tupin, and Xiaoxiang Zhu. Fusion of sar and optical remote sensing data — challenges and recent trends. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5458–5461, 2017. 1

[52] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 153–160, 2019. 4, 5, 7, 14

[53] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. 3, 5

[54] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts, 2023. 3

[55] Yongwei Sheng and Zong-Guo Xia. A comprehensive evaluation of filters for radar speckle suppression. *IGARSS '96. 1996 International Geoscience and Remote Sensing Symposium*, 3: 1559–1561 vol.3, 1996. 5

[56] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 5, 6, 7, 13

[57] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2022. 1, 2

[58] Melanie K. Vanderhoof, Laurie Alexander, Jay Christensen, Kylen Solvik, Peter Nieuwlandt, and Mallory Sagehorn. High-frequency time series comparison of sentinel-1 and sentinel-2 satellites for mapping open and vegetated water across the united states (2017–2021). *Remote Sensing of Environment*, 288:113498, 2023. 1

[59] Stefano Vincenzi, Angelo Porrello, Pietro Buzzega, Marco Cipriano, Pietro Fronte, Roberto Cuccu, Carla Ippoliti, Annamaria Conte, and Simone Calderara. The color out of space: learning self-supervised representations for earth observation imagery. *CoRR*, abs/2006.12119, 2020. 2

[60] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1, 2

[61] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards remote sensing foundation model, 2022. 2, 7

[62] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops, 2023. 5, 6, 12

[63] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *CoRR*, abs/2111.09886, 2021. 2, 3, 4, 5, 12

[64] Chi Xu, Yanling Ding, Xingming Zheng, Yeqiao Wang, Rui Zhang, Hongyan Zhang, Zewen Dai, and Qiaoyun Xie. A comprehensive comparison of machine learning and feature selection methods for maize biomass estimation using sentinel-1 sar, sentinel-2 vegetation indices, and biophysical variables. *Remote Sensing*, 14(16), 2022. 1

[65] Fang Xu, Yilei Shi, Patrick Ebel, Lei Yu, Gui-Song Xia, Wen Yang, and Xiao Xiang Zhu. Glf-cr: Sar-enhanced cloud removal with global–local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:268–278, 2022. 1, 6, 7

[66] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*, 2022. 3, 8

A. Optimization on Pretraining Data

A.1. Performance with other choices of pretraining data.

To optimize GeoPile-2, we initially focused on optimizing GeoPile-2-RGB. As previous research has indicated, a successful pretraining dataset requires rigorous testing of each component [43]. Thus, we conducted a series of experiments on each individual dataset. These experiments involved the use of ImageNet [17] with 3 million images, GeoLifeCLEF with 3.3 million images, and the Functional Map of the World (FMoW) [14]. For FMoW, we segmented the dataset into tiles of size 384, leading to a total of 6 million images. This diverse selection of datasets allowed us to comprehensively test and optimize our pretraining approach for GeoPile-2.

To ensure other variables, such as the backbone architecture and pretraining methodologies, do not skew our results, we chose to employ the Swin-base [35] and committed to pretraining from scratch. In line with our aim for equitable comparison, we also adhered to the same seven downstream tasks as delineated in the previous report [39]. The results are shown in Table 7 and Table 8. This approach creates a consistent testing environment across all datasets, reducing the potential for bias or error.

Interestingly, upon integrating the GeoLifeCLEF into our testing framework, we observed a downturn in performance on downstream tasks. This result signifies that not all datasets necessarily contribute to improved model performance, and their selection demands careful consideration.

Even though the addition of both the Functional Map of the World dataset and ImageNet gave rise to performance metrics that were commensurate with those achieved by GeoPile-2-RGB, these new dataset additions were not as efficient. The key reason for this inefficiency was the significantly larger size of the pretraining dataset, which introduced higher computational costs and longer processing times. This finding highlights the importance of carefully balancing dataset size and complexity with computational efficiency in the model training process.

A.2. Performance without RGB modality

RGB modalities are singled out because of the abundance of RGB datasets that come from various sources beyond just Sentinel-2. For instance, MillionAID [37], a dataset comprised of a wide range of RGB images, is sourced from multiple satellites, including GeoEye, WorldView, QuickBird, IKONOS, and SPOT satellites, among others. Additionally, a previous study [15] found that using only Sentinel-2 data for pretraining does not yield optimal performance in the downstream evaluation. Therefore, we sought to diversify our sources and include a wider range of RGB images in our pretraining data. This breadth of data sources significantly enriches the diversity of the RGB modality in our study.

Despite overlapping GSD in some RGB modality, more geospatial features will be included. Although these datasets may not provide an imaging spectrum as wide as Sentinel-2, they enhance the entropy of pre-training data, which has been proven to be effective in [39], which is demonstrated by Table 9.

B. Pretraining Details

B.1. Pretraining Settings

Masking. All hyper-parameters are listed in Table 10. We implement a masking strategy that maintains consistency around different channels within the same sensor, applying the mask at the same locations. However, when it comes to different sensors, we employ a varying masking approach, ensuring that the mask is applied at different locations. This methodology allows us to preserve sensor-specific information while investigating inter-sensor discrepancies effectively.

Heterogeneous batch size. Given the disparity in the number of images obtained from different sensors, we employ a heterogeneous batch size strategy for our training process. This methodology adjusts the batch size in proportion to the amount of data sourced from each individual sensor. In essence, during each epoch of our training process, every type of sensor is iterated through once, irrespective of the data volume associated with that particular sensor. This ensures that all sensor types have an equal chance to contribute to the model's learning process, fostering a more balanced and comprehensive training regimen. Alongside this, we also adjust the learning rate proportionally in accordance with the batch size allocated per sensor.

C. Downstream Experiments

C.1. Model size

Regarding the number of parameters, we followed a standard backbone for pretraining, the details of which have been reported in [63]. Comparisons between training from scratch and using ImageNet pretrained weights have been provided in Table 11 and corroborated by previous studies [8, 15, 38, 39, 62].

C.2. Experimental settings

There are primarily two ways to leverage pretrained weights, as depicted in Figure 6. The first approach involves feeding each sensor through encoders that share weights. The resulting embeddings are then concatenated and fed into the classifier. In the second approach, all sensor data are stacked together in the color channel prior to patchification. This approach resembles the multiMAE method [5], where the projected patches from all modalities are concatenated

Dataset	# Image	OSCD (F1)	DSFIN (F1)	BEN 10%	BEN 1%
GeoPile [39]	600K	**57.5**	66.2	86.4	79.3
GeoPile-2-RGB	1.7M	57.1	**70.4**	**86.8**	**79.6**
GeoPile-2-RGB + ImageNet [17]	3M	**57.5**	69.2	86.4	79.5
GeoPile-2-RGB + GeoLifeCLEF	3.3M	56.1	61.6	86.1	78.9
GeoPile-2-RGB + FMoW [14]	6M	58.2	69.3	86.2	79.1

Table 7. Results of downstream tasks with different pretraining datasets: change detection and classification

Dataset	# Image	WHU	Vai.	SN2 (PSNR)	SN2 (SSIM)
GeoPile [39]	600K	90.1	75.1	**22.626**	0.645
GeoPile-2-RGB	1.7M	**90.6**	75.9	22.599	**0.658**
GeoPile-2-RGB + ImageNet [17]	3M	90.5	**76.1**	22.107	0.631
GeoPile-2-RGB + GeoLifeCLEF	3.3M	89.1	74	16.663	0.512
GeoPile-2-RGB + FMoW [14]	6M	90.2	75.7	22.448	0.638

Table 8. Results of downstream tasks with different pretraining datasets: segmentation and super-resolution

Pretraining sensor modality	10% BEN	cloud removal
Metric	mAP (↑)	SAM (↓))
SAR (in Figure 3)	84.2	8.37 ± 0.057
Sentinel-2 (in Figure 3)	86.6	8.33 ± 0.034
RGB	86.4	10.45 ± 0.12
w/o RGB	86.6	9.67 ± 0.12
GeoPile-2	**87.7**	**7.51** ± 0.057

Table 9. Results pretrained with single modality or without RGB modality.

into a single sequence. Our experiments on 1% of the BEN dataset [56], listed in Table 12, demonstrate that both methods yield comparable results. However, the latter approach is more computationally efficient, meaning that the former approach takes longer time to reach the optimal performance and consumes more memory as well. Therefore, all results mentioned in the main text utilize this second approach. Importantly, in both cases, no masking is performed during the transfer phase.

C.3. Visualization

We present some quantitative results of segmentation in Figure 7 respectively.

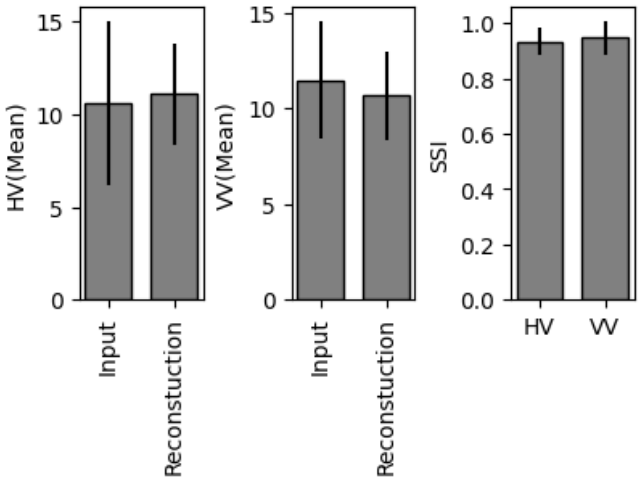

Figure 5. Figure R1: SAR backscatter statistics comparing input and reconstruction using the MIM. The two bands of SAR are HV and VV. The mean and standard deviation for the HV band are shown on the left, while those for the VV band are displayed on the right. The Speckle Suppression Index (SSI) values are presented in the right panel. An SSI value closer to one indicates that the mean and standard deviation remain consistent before and after reconstruction.

Hyper-parameter	Value
Image size	192×192
Optimizer	AdamW
β_1	0.9
β_2	0.999
Eps	1.0×10^{-8}
Momentum	0.9
Weight decay	0.05
Learning rate	$\{1.0 \times 10^{-4}, 0.25 \times 10^{-4}, 1.0 \times 10^{-5}\}$ for RGB, Sen12MS [52] and MDAS [29]
Warm up learning rate	5.0×10^{-7}
Weight decay	10^{-5}
Batch size	{128, 32, 12} per GPU for RGB, Sen12MS [52] and MDAS [29]
Training epochs	800 or 100
Warm up epochs	10
Learning rate decay	Multistep
Gamma	0.1
Multisteps	[700,] for 800 or [] for 100
# Experts	8
MoE blocks	1, 3, 5, 7, 9, 11, 13, 15, 17 (Every other swin block)
Top-value (k)	1
Capacity factor	1.25
Aux loss weight (λ)	0.01
Mask patch size	32
Mask ratio	0.6

Table 10. Hyperparameters of msGFM pretraining.

Model	SeCo	SatMAE	MoCoV2	DINO-MC	GFM	msGFM
# of trainable parameters	23M	307M	23M	48.6M	89M	89M

Table 11. Model size

Finetuning Method	BEN 1%
1	80.8
2	80.8

Table 12. Results of BEN when comparing different downstream transfer methods illustrated in Figure 6

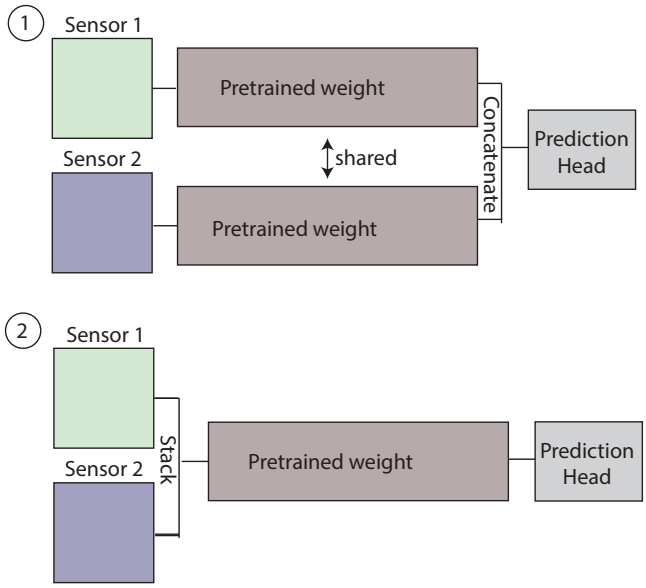

Figure 6. Two methods of downstream transfer. In the top panel, every sensor is fed into a separate encoder initialized with msGFM pretrained weight. The embeddings from the last layer are concatenated, and then fed through the prediction head, such as classifier and segmentation decoder. In lower panel, images are concatenated along the color channel and then fed through one encoder initialized with msGFM pretrained weight.

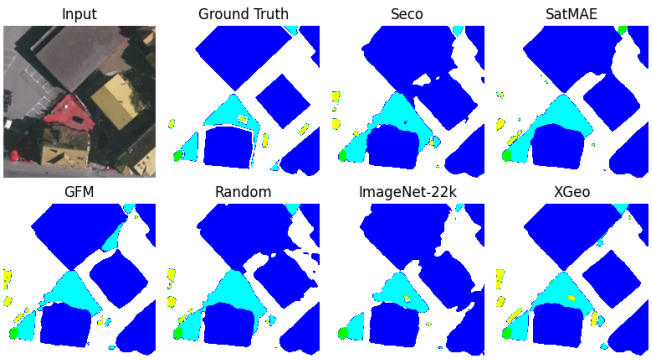

Figure 7. A display of qualitative results showcasing segmentation outcomes from msGFM in comparison to other competitive methods.