Enhancing Remote Sensing Representations Through Mixed-Modality Masked Autoencoding

Ori Linial, George Leifman, Yochai Blau, Nadav Sherman,
Yotam Gigi, Wojciech Sirko, Genady Beryozkin

Google Research

Abstract

This paper presents an innovative approach to pretraining models for remote sensing by integrating optical and SAR (Synthetic Aperture Radar) data from Sentinel-2 and Sentinel-1 satellites. Using a novel variation on the masked autoencoder (MAE) framework, our model incorporates a dual-task setup: reconstructing masked Sentinel-2 images and predicting corresponding Sentinel-1 images. This multi-task design enables the encoder to capture both spectral and structural features across diverse environmental conditions. Additionally, we introduce a "mixing" strategy in the pretraining phase, combining patches from both image sources, which mitigates spatial misalignment errors and enhances model robustness. Evaluation on segmentation and classification tasks, including Sen1Floods11, BigEarthNet, and UrbanSRSeg8, demonstrates significant improvements in model performance and generalizability across diverse remote sensing applications.

1. Introduction

Self-supervised learning has recently transformed computer vision, allowing models to extract meaningful representations from large-scale image datasets without the need for labels [11, 12, 20, 22, 23, 27, 29, 43]. This approach has shown strong performance in all tasks, including classification [9, 11, 23, 34] and segmentation [23, 47]. Self-supervised methods are particularly valuable in fields where unlabeled data is abundant, but creating labeled datasets is challenging and costly due to the specialized expertise required. Remote sensing is one such field where self-supervised methods hold immense promise due to the complexity of annotating data. Although imagery from sources such as satellites, aerial sensors, and other remote systems is readily available, obtaining labeled data for tasks such as segmentation or classification demands considerable additional effort and resources.

Remote sensing plays a critical role in applications with far-reaching societal impact, including climate monitoring, poverty mapping, disaster response, agriculture, and urban planning [2, 4, 8, 28, 33, 38, 45, 48]. Unlike many other image domains, where labeling can be crowd-sourced or automated, remote sensing data often requires extensive domain expertise to interpret, especially when identifying land cover types, monitoring environmental changes, or tracking natural hazards. Although initiatives are underway to create labeled remote sensing datasets, such as Sen1Floods11 [7] and BigEarthNet [41], the complexity and diversity of data (from multispectral satellite imagery to SAR scans) make comprehensive annotation difficult and time-consuming. As a result, large portions of remote sensing data remain unlabeled, creating a significant challenge for developing accurate, real-world models. Self-supervised learning offers a promising solution, encouraging its adoption in the remote sensing domain, where it can leverage abundant unlabeled data to generate meaningful representations for critical applications [15, 21].

In recent years, a variety of self-supervised learning methods have been successfully adapted to the remote sensing domain, including masked autoencoders (MAE) [15,37] and contrastive learning frameworks [13, 21, 25, 26, 46]. Typically, these models process a large amount of image data acquired by a single remote sensing source and learn a meaningful semantic representation of the given image. For example, SatMAE [15] trains on images taken from the Sentinel-2 satellite or on high-resolution RGB images.

Each remote sensing source presents unique advantages and challenges. Sentinel-2, for instance, provides low-resolution multispectral imagery, which is valuable for analyzing vegetation health, soil properties, and water bodies but may lack fine spatial details in comparison to high-resolution RGB sources. On the other hand, Sentinel-1 captures radar images, offering all-weather, day-and-night imaging capabilities that are essential for monitoring surface changes, even under cloud cover. Integrating these data sources, essentially treating them as different modalities, is

not straightforward. However, combining multiple modalities has the potential to improve model robustness and scene understanding by leveraging the complementary strengths of each data source [10]. We demonstrate some of the differences between Sentinel-1 and Sentinel-2 in Figure 1.

Recent work has been trying to integrate several remote sensing modalities such as Sentinel-2, Sentinel-1 or even high-resolution imagery [13, 18, 21, 25, 26, 46], focusing mainly on the well known contrastive framework. This framework focuses on maximizing mutual information between representations from different modalities or views, aiming to make similar representations closer and dissimilar ones farther apart. This approach can capture shared information but may overlook unique, complementary features specific to each modality. In contrast, an MAE approach–by masking and reconstructing parts of each modality–can potentially learn a richer, more comprehensive understanding of each modality's unique characteristics as well as their interdependencies [5]. This capability is especially valuable when modalities have complementary information that may not be fully leveraged through contrastive learning alone.

In this work, we provide two methodologies for integrating different remote sensing data sources:

1. **AuxMAE.** This model is based on the standard MAE, where images from a single source serves as the model's input, and the model's goal is to both reconstruct the input image as in standard MAE, but also to predict an auxiliary image taken from a different source. For example, the input could be images taken by Sentinel-2, and the output is both the input image and an image from Sentinel-1.

2. **MixMAE.** This model takes the standard MAE, and instead of using masks over the input images, it mixes patches from all sources to a single image. The mixture image is then encoded and decoded back to both input modalities to predict the masked patches.

We assess the models capabilities by testing them on three different downstream tasks, and comparing to an MAE trained on a single source. The downstream tasks are the well-known Sen1Floods11 [7] and BigEarthNet [41], and an additional, non-public, segmentation task, UrbanSRSeg8. Our tests show that both AuxMAE and MixMAE are superior to the baseline model not only when finetuning the entire model, but also when freezing the encoder, a very desired scenario due to lower computation cost.

2. Related Work

2.1. Self-Supervised Learning in Remote Sensing

Self-supervised learning (SSL) has emerged as a powerful tool in fields where labeled data is limited, enabling models to learn useful representations directly from raw, unlabeled data [11, 12, 20, 23, 43]. By predicting parts of the data or creating pseudo-labels from inherent data patterns, SSL reduces the reliance on manual annotation while often achieving performance comparable to supervised methods. This ability to unlock patterns in large datasets without labeled input makes SSL particularly valuable for remote sensing, where abundant but unlabeled imagery from satellites and aerial sensors is available.

In remote sensing, SSL methods have traditionally focused on single-source data, where models learn to generate representations for individual data types, such as Sentinel-2 [15, 32], Sentinel-1 [26], or high resolution RGB images [3, 15, 30, 37, 40, 42], to name but a few.

In addition, some studies are also exploring the inclusion of external knowledge, such as geographical or topographical information, to further improve model performance. For example, Li et al. [31] incorporated geographical knowledge into a remote sensing model, allowing for more accurate predictions by accounting for spatial relationships and contextual information. This approach is part of a growing trend to fuse domain-specific knowledge with raw image data, enhancing the model's understanding and enabling more precise, context-aware decision-making across tasks like disaster response and land-use planning.

2.2. Multi-Modal Self-Supervised Learning in Remote Sensing

Multi-modal SSL has become an important area of study in computer vision, where combining information from multiple data modalities can yield richer and more robust representations. By integrating data sources with complementary properties–such as RGB and depth images, visual and textual data, or even visual and audio data–multi-source SSL has shown significant potential to improve model generalization across various tasks, including classification, segmentation, and retrieval [5, 6, 17, 19, 36]. Techniques such as contrastive learning and masked prediction objectives are commonly employed to align and fuse features across modalities, enhancing the model's understanding of complex scenes.

In remote sensing, multimodal SSL approaches are particularly valuable for applications where diverse data sources, such as multispectral (e.g., Sentinel-2) and radar imagery (e.g., Sentinel-1), provide complementary insights. Unlike single-source models, which are constrained by the limitations of individual sensors, multi-source models can leverage the unique strengths of each modality to develop more robust and generalizable representations. Most recent multimodal SSL work in remote sensing has focused on contrastive learning for pretraining the model [13, 21, 25, 26, 46] or training a teacher-student model [18], leaving MAE approaches largely underexplored for integrating multiple

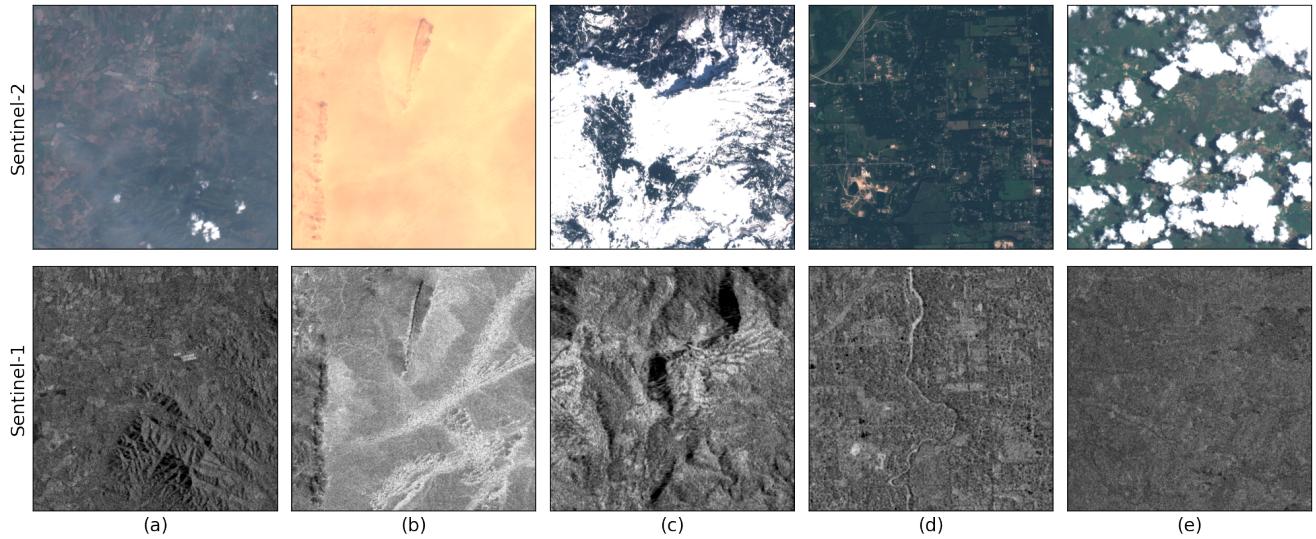

Figure 1. Comparing the Sentinel-2 and Sentinel-1 sources. Sentinel-2 (top row) provides low-resolution multispectral imagery, and Sentinel-1 (bottom row) captures radar images, offering all-weather, day-and-night imaging capabilities. Their differences can be seen clearly in this figure. (a) and (e) demonstrate that some images could be slightly to heavily cloudy, making Sentinel-2 unusable for these images. (b) demonstrates that in sandy areas understanding the details in Sentinel-2 image is almost impossible. (c) shows the same for snowy regions, where the fine details in the Sentinel-2 image are not observable easily. Finally, (d) demonstrates that some information in the image is more apparent in Sentinel-1 like some rivers and roads. These examples serve as motivation for incorporating both sources in our model.

modalities. Inspired by the success of MultiMAE in multi-modal learning for other computer vision tasks [5] and by the capabilities of SatMAE [15] and Presto [44] as MAE models in the remote sensing domain for a single modality, we argue that MAE architectures are also well-suited to multi-source SSL in remote sensing. This paper explores two primary multimodal MAE-based architectures, aiming to capture both shared and complementary features for a more nuanced, holistic representation of multimodal satellite data.

Building on this idea, recent work has begun to explore variations of masked autoencoders in multi-modal settings. For instance, the Multi-Pretext Masked Autoencoder proposed in [35] provides an innovative perspective on integrating diverse modalities. Another study [14] introduced an incomplete multimodal learning approach, combining reconstruction and contrastive loss. While this method is similar in spirit to our MixMAE pretraining strategy, it uses relatively small datasets for pretraining and thus it was hard to see the full potential of the approach.

3. Model

In this section we describe on the pretraining of both our models AuxMAE and MixMAE. We start by providing details on the image acquisition process and how the pretraining datasets were generated. We then describe both models, giving details on the architectures used, the training procedure, and details regarding finetuning the models for downstream tasks.

3.1. Acquisition of pretraining data

For the pretraining phase, we constructed a remote sensing dataset consisting of images acquired from two distinct satellite sources: Sentinel-1 and Sentinel-2. The Sentinel-2 images are collected at top-of-atmosphere (L1C) level and comprise 13 spectral bands. In contrast, the Sentinel-1 images are synthetic aperture radar (SAR) images, and were acquired using the Interferometric Wide (IW) swath mode with two polarization channels: VV and VH. We note that Sentinel-1 operates in several acquisition modes, each tailored for specific applications (e.g., Stripmap, Extra-Wide Swath, and Wave modes)[1]. The key difference between the Sentinel-1 and Sentinel-2 sources lies in their sensing capabilities; Sentinel-2 captures detailed optical imagery across multiple bands that allows for more accurate discrimination of different land cover types, while Sentinel-1 provides capability to capture data under all weather conditions and during both day and night.

The dataset was formed by coupling geo-spatially and temporally aligned Sentinel-1 and Sentinel-2 images. The first stage involved collecting Sentinel-2 images taken from

[1]More information regarding Sentinel-1 acquisition modes and applications at https://sentiwiki.copernicus.eu/web/s1-applications

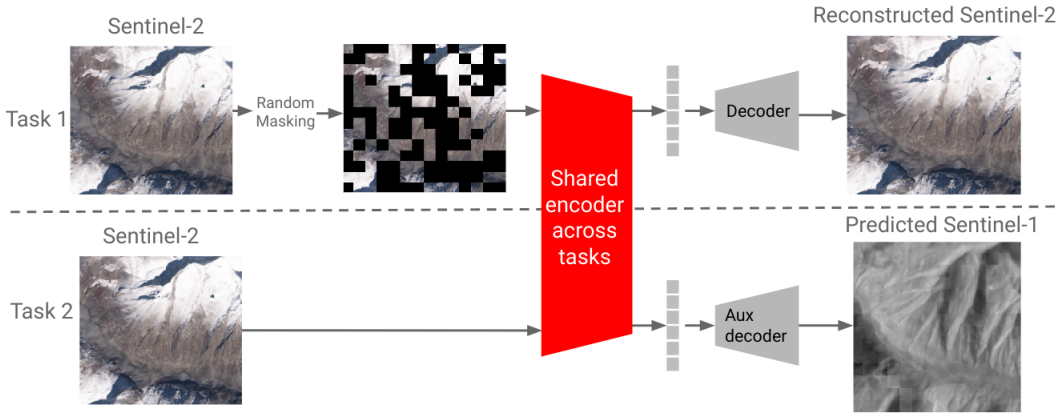

Figure 2. Pretraining the AuxMAE model. The model consists of a standard MAE, and an additional aux decoder for predicting a Sentinel-1 image given an unmasked Sentinel-2. Both tasks share the same encoder, and are trained in an AGD fashion. Meaning, at each time step the model optimizes for a single task. The loss for task 1 is an MSE loss on the unseen Sentinel-2 patches, and the loss for task 2, is the MSE loss on the entire Sentinel-1 predicted image.

all six continents across different seasons in the years 2022 and 2023. Subsequently, the alignment process involves the following steps for each Sentinel-2 image: (1) identifying all corresponding Sentinel-1 images based on matching latitude and longitude coordinates, and (2) selecting the closest Sentinel-1 image by acquisition time. Sentinel-1 images taken more than one week from the corresponding Sentinel-2 image are excluded from the dataset, as are Sentinel-2 images without a corresponding Sentinel-1 image. This rigorous alignment process has resulted in a total of approximately 40 million pairs of images.

This combined dataset is well-suited for pretraining, as Sentinel-2's rich spectral data supports the learning of spatial and spectral patterns critical for accurate land cover analysis. By incorporating Sentinel-1 as an auxiliary target, we leverage SAR's complementary ability to capture structural features regardless of weather or lighting conditions. Nevertheless, this process can be extended to incorporate additional remote sensing sources, such as hyperspectral satellites with hundreds of bands or high-resolution RGB imagery for urban areas (though this is out of scope for this paper). The only requirement is that the images share the same spatial coordinates and approximate acquisition times. A significant challenge when integrating multi-source data is alignment: all sources must provide data for the same spatial and temporal regions; otherwise, unaligned images are excluded from the dataset. For example, if a high-resolution RGB source is only available for urban areas, relying on this data would restrict the usable coverage of Sentinel-2 and Sentinel-1 to those urban areas as well, leading to significant data loss. To overcome this issue, we applied Alternating Gradient Descent (AGD) [1], as described in Section 3.2. AGD enables training with multiple datasets, each corresponding to different source combinations—such as one dataset combining Sentinel-2 and Sentinel-1, and another combining Sentinel-2 with high-resolution imagery. During training, only one dataset is used at a time, allowing the model to learn from each source combination without imposing strict alignment requirements across all sources. This approach offers greater flexibility and reduces data loss, ensuring broader and more effective utilization of diverse remote sensing data.

3.2. Pretraining AuxMAE

The AuxMAE model is based on the widely-used masked autoencoder (MAE) [22] algorithm, a self-supervised framework specifically designed for vision tasks. The core concept of the MAE model is to randomly mask patches of the input image, with the model's objective being to reconstruct the missing patches at the output. Typically, MAE is first pretrained on a very large amount of unannotated data, and then its backbone is finetuned for downstream tasks using a smaller set of annotated data. In the context of remote sensing imagery, earlier works have utilized MAE for the remote sensing tasks, such as using Sentinel-2 images as the input images, and reconstructing them at the output [15].

In this work, we build upon the MAE-based model, and test whether adding an auxiliary task to the original model allows the model to achieve better results on downstream tasks. In other words, we examine whether adding a Sentinel-2-to-Sentinel-1 objective enhances the encoder's semantic understanding of the input images. We stress that the model still has the original goal of reconstructing the observed masked Sentinel-2 image. To that end, we add an additional decoder (namely, aux-decoder) with a goal of outputting a prediction of the Sentinel-1 image, given the original Sentinel-2 image.

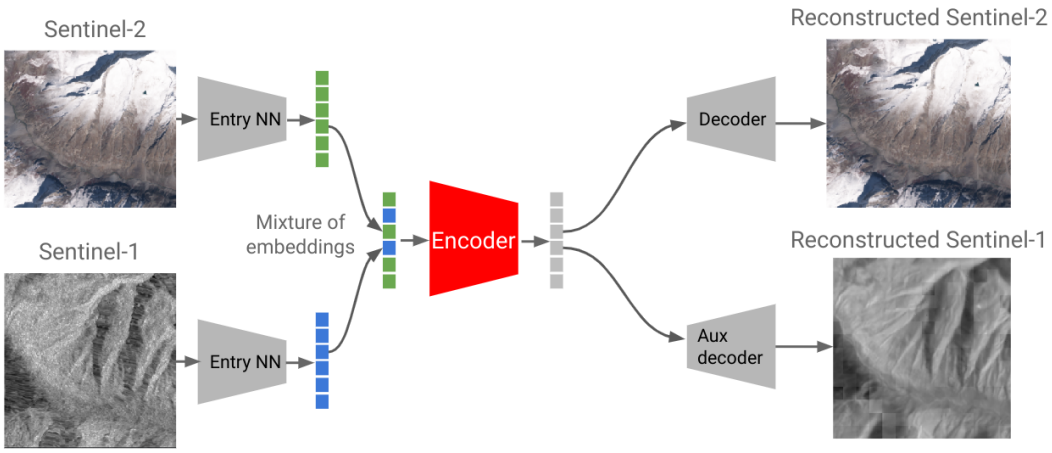

Figure 3. Pretraining the MixMAE model. The model takes images from two different sources, inputs each image to a different entry NN, and obtains two different embedding sequences representing the given images. Then, each embedding in the final embedding sequence is taken randomly from one of the two embedding sequences. This mixture of embeddings then serves as input for the encoder. The encoder's output is then used for two decoders to reconstruct the both given images.

To train the system, we used alternating gradient descent (AGD) [1, 24], alternating each batch between one of the following tasks: (1) reconstruction of the given Sentinel-2 image or (2) prediction of the Sentinel-1 image. For the first task (Sentinel-2 to Sentinel-2), we use the MAE in the most standard way: a random mask is applied to the Sentinel-2 image, the masked image is fed into the shared encoder, the encoder's output is passed to the decoder, and finally, the MSE loss is evaluated between the decoder's output and the unmasked Sentinel-2 image. For the second task (Sentinel-2 to Sentinel-1), the Sentinel-2 image is fed into the shared encoder (without masking), the encoder's output is passed to the auxiliary decoder, and the MSE loss is evaluated between the decoder's output and the Sentinel-1 image. The use of a single shared encoder combined with two distinct decoders ensures that the model can handle variations in the spatial resolution of the input and target data sources. For example, one could use a Sentinel-2 image with a spatial resolution of 10m per pixel as input and have an auxiliary task of predicting an image with a resolution of 1m per pixel (alongside the reconstruction task). The general process is described in Figure 2.

According to Jain et al. [25] (and further discussed by Akbary et al. [1]), AGD with two alternating tasks has been proven to have the same effect as standard GD with a single task that balances between two loss terms, as long as each task has a convex optimization step. However, AGD offers a distinct advantage in our setup, as it allows us to effectively utilize both masked and unmasked versions of the Sentinel-2 input images. Additionally, aligning images from two different sources spatially and temporally can be challenging, often resulting in a dataset biased towards specific regions or times. For example, if the auxiliary data source tends to capture images more frequently in urban areas, then the intersecting dataset of aligned image pairs will have a higher concentration of urban scenes. This can limit the model's generalizability across different environments.

AGD addresses this problem by using the complete dataset of Sentinel-2 input images, consisting of approximately 40M Sentinel-2 images, alongside the intersecting dataset of paired images from both sources, rather than relying solely on the intersecting set. Moreover, in some cases, the auxiliary dataset is much smaller than the primary input data, making AGD even more suitable, since it maximizes the available data for pretraining. Thus, AGD not only mitigates regional biases in the intersecting dataset but also leverages a larger, more diverse set of images to enhance the model's adaptability across varied downstream tasks.

3.3. Pretraining MixMAE

The main goal of this paper is to suggest a method of incorporating additional data sources to improve the pretraining phase of a model, to make it more resilient, robust, and have better generalization. To that end, we evaluated an additional strategy for model pretraining, where both data sources are mixed at the encoder's input. The first stage in this model is to pass both the original input image (Sentinel-2) and the auxiliary image (Sentinel-1) through the entry layers of the encoder, resulting in two sequences of L embeddings (L is determined by the image dimensions, patch size, and so on), where each embedding corresponds to a single patch in the input image. After obtaining an embedding for each image, we create the total embedding randomly sampling $0 < k < L$ embeddings from the first sequence and the remaining $L - k$ embeddings from the second sequence, for some predetermined k (we use $k = L/2$).

The result embedding sequence is a random mix between the two embedding sequence and is used as an input to the encoder's transformer. The mixing strategy basically mimics the behavior of the standard MAE, with the main difference that instead of having masked patches, we have patches from a different source. Subsequently, the output of the encoder serves as input to two decoders, one for each source. In our case, the input image is a mixture of Sentinel-2 patches and Sentinel-1 patches, and the decoders' goal is to reconstruct the missing patches from both images. The loss function is then evaluated only on the unseen patches as a standard MSE for each source, and the final model loss is the combination of them:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{Sentinel-1}} + (1-\lambda)\mathcal{L}_{\text{Sentinel-2}} \quad (1)$$

where $\mathcal{L}_{Sentinel-1}$ is the loss between unseen Sentinel-1 patches and the ground truth, and $\mathcal{L}_{Sentinel-2}$ is the loss between unseen Sentinel-2 patches and the ground truth. λ is a predetermined hyperparameter we set to 0.5. The general process is described in Figure 3, and in short in Figure 4.

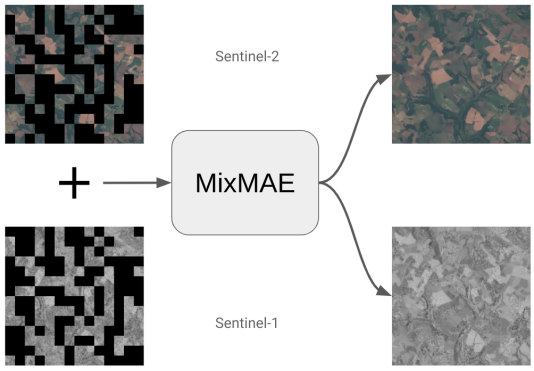

Figure 4. An illustration of the pretraining of MixMAE. The model first combines two masked images from two sources with complementing masks (i.e., each patch is taken either from the first source or from the second at random), and then outputs the two reconstructed images. The images in this figure were indeed taken from the reconstruction of the MixMAE model while pretraining.

Although Sentinel-1 and Sentinel-2 images are generally aligned, minor misalignments can lead our original model to produce blurry Sentinel-1 outputs, potentially degrading the quality of encoder features. In the mixing approach, the model should have the ability to correct spatial alignment errors between the two images by leveraging patches from both images. The alignment problem can have crucial effect in tasks where the spatial resolution is very different. For example, when one source is Sentinel-2 (with 10m per pixel) and the other one is a high-res source (with 0.5m per pixel). In this case, even small misalignments between the sources can interfere with the models ability to learn. We show in Section 4 that although Sentinel-2 and Sentinel-1 are generally spatially aligned (by longtitude and latitude), the mixing model obtains better results in all downstream tasks.

On the other hand, MixMAE does not enjoy the computational advantages of the MAE since the sequence in the input of the encoder's transformer contains embeddings of all patches, where in standard MAE the input of the transformer is just the visible patches (usually, 25% of the entire patches).

4. Experiments

In this section, we evaluate both our models ability to generalize across diverse remote sensing tasks, and provide details on the tasks, the pretraining of the models and the finetunining for the downstream tasks, and provide the results.

4.1. Downstream tasks

The primary goal of the pretrained models is to enable rapid adaptation to new tasks, with a small amount of data, few training steps, and ideally, minimal fine-tuning of the full model. To achieve this, we followed standard practice by using the trained encoder and attaching a new prediction head with a much smaller parameter count for each downstream task. For each task, we added the required prediction head (segmentation or classification) and evaluated the corresponding loss.

To assess our models' adaptability to new remote sensing datasets, we selected two segmentation tasks–Sen1Floods11 [7] and UrbanSRSeg8–and one classification task, BigEarthNet [41]. We stress that our goal is not necessarily to achieve state-of-the-art results on all tasks, but to isolate the effect of having additional source of information on an MAE based pretrained model. However, it is noteworthy to mention that our results are comparable to the SOTA results on these benchmarks, while using only ViT-Base encoder which is lighter than other baselines which usually use ViT-Large. For example, in BigEarthNet, we achieve 91.56% using ViT-Base, while SatMAE [15] achieves 92.09% using Vit-Large.

4.1.1 Sen1Floods11

Sen1Floods11 is a flood-detection segmentation task using Sentinel-2 images, with a total of 252 training images and 90 test images. An important note regarding Sen1Floods11 is that we use only the Sentinel-2 images it provides, and dropped the Sentinel-1 images. We did this so that our pretraining models, which used Sentinel-1 while pretraining, will not have an advantage over the baseline model that trained only on Sentinel-2. The model trains to predict the three provided classes: background, water, and clouds. However, to evaluate our models, we used the common

	Sen1Floods11	BigEarthNet	UrbanSRSeg8
Task	Segmentation	Classification	Super resolution segmentation
Source	Sentinel-2	Sentinel-2	Sentinel-2
Bands	13	13	13
Resolution	10m	10m	10m
Image Size	512×512	120×120	256×256
Label Classes	2 (existance of water)	19 (see Section 4.1.3)	8 (building, paved road, unpaved road, ground, parking lot, tree, water, other human made)
Label Size	512×512	-	1024×1024
Train Images	252	311,667	33,899
Test Images	90	2,432	103,728
Train Regions	11 flooding events from 6 continents	Europe: Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland	Global

Table 1. Benchmarks details

intersection-over-union (IoU) metric achieved on the water class over the test set, as suggested in other papers [26].

4.1.2 UrbanSRSeg8

UrbanSRSeg8, a non-public dataset, is a super-resolution full-scene semantic segmentation task with eight possible labels (e.g., building, road, tree, water). The input is a Sentinel-2 top-of-atmosphere (L1C) images of size 256×256. The label is a 1024×1024 semantic segmentation of that same area in resolution of 2.5m, derived from high-resolution imagery. It contains 33,899 training images and 2,432 test images. The metric we used for this task is the mean IoU over all classes, which is calculated by computing the IoU per class over all images, and then averaging over all classes.

4.1.3 BigEarthNet

BigEearthNet is a multi-label land-cover scene classification task with 19 possible labels, from coastal wetlands to pastures and forests[2]. The dataset is composed of 311,667 training images and 103,728 test images. BigEarthNet images are derived from Sentinel-2 and correspond to surface reflectance (L2A) data, making it distinct from the top-of-atmosphere (TOA, L1C) data used during pretraining. For this experiment, we finetuned our pretrained TOA model to adapt it to the surface reflectance domain. The metric we used for this task is the mean average precision (mAP).

[2]The 19 classes as suggested by Sumbul et al. [41] are: Urban fabric, industrial or commercial units, arable land, permanent crops, pastures, complex cultivation patterns, land principally occupied by agriculture with significant areas of natural vegetation, agro-forestry areas, broad-leaved forest, coniferous forest, mixed forest, natural grassland and sparsely vegetated areas, moors, heathland and sclerophyllous vegetation, transitional woodland-shrub, beaches, dunes, sands, inland wetlands, coastal wetlands, inland waters, marine waters.

4.2. Training details

We trained both our models on a total of approximately 41M couples of Sentinel-2 and Sentinel-1 images, for 1M train steps with AGD training strategy. This effectively means that the model optimizes standard MAE for the Sentinel-2 to Sentinel-2 task for 500K steps, and optimizes the Sentinel-2 to Sentinel-1 prediction task for the another 500K steps, alternating between the tasks at each step. The encoder in all experiments has a ViT-base architecture [16] and with a patch size of 16×16 pixels. The decoder (for reconstructing Sentiel-2 images) and aux-decoder (for predicting Sentinel-1 images) we used for all experiments have the same ViT-Lite architecture with approximately 21M parameters.

For the mixing model, we trained the model for 500K steps with the same architecture as our standard MAE model with a single ViT-Base encoder, and two ViT-Lite decoders.

For the downstream tasks we used the pretrained encoder and added a new prediction head for each task, where for the segmentation downstream tasks we added used the ViT-Lite architecture and an additional segmentation head, and for classification we used linear probing. We trained UrbanSRSeg8 for 10K steps, Sen1Floods11 for 30K steps, and BigEarthNet for 30K steps.

For all training procedures we used a batch size of 4096 and the AdaFactor optimizer [39].

4.3. Results

We tested both our pretrained models on all three downstream tasks in two scenarios: (1) Finetuning the encoder

and optimizing the prediction head (namely, finetune), and (2) optimizing the prediction head alone (namely, frozen). The reason we evaluated both scenarios is that each tests a distinct aspect of the model's adaptability. The finetuning scenario assesses the model's ability to optimize both the encoder and prediction head together, allowing it to fully adapt to each task's specific characteristics, which can reveal the model's maximum potential performance. On the other hand, the frozen scenario tests how well the model performs under lower-compute settings by using only the prediction head, which is important for applications where computational resources are limited. This setup evaluates the encoder's effectiveness at learning general, transferable representations during pretraining that can support new tasks with minimal additional tuning. We also tested the mixing model on the finetune scenario, and added a baseline of using an MAE that was trained only on the Sentinel-2 images, without any access to other sources such as Sentinel-1. We summarize all results in Table 2.

Comparing the results on the different tasks, scenarios, and models, we observe first that our AuxMAE has significantly better results than the Sentinel-2 MAE baseline for the frozen scenario, suggesting that the encoder has indeed learned a better semantic understanding. This result is especially important in applications that require robust performance under limited computational resources, as it indicates that AuxMAE can provide strong task-specific representations even when only the prediction head is optimized. Such efficiency is crucial for remote sensing applications deployed in resource-constrained environments, where quick adaptation to new tasks without extensive retraining is highly valued.

On the other hand, in the finetune scenario, we observe that the results of the AuxMAE are better in Sen1Floods11, and in BigEarthNet, but not in UrbanSRSeg8. We believe that the reason for this is that UrbanSRSeg8 is a relatively big dataset (compared to Sen1Floods11), suggesting that finetuning on a larger dataset like UrbanSRSeg8 allows the model to adapt effectively without the added benefit of Sentinel-1 data. This may indicate that the additional semantic information from Sentinel-1 is more impactful in cases where the task dataset is smaller and benefits more from enriched pretraining. In contrast, for larger datasets, the model can potentially learn task-specific representations during finetuning, diminishing the reliance on pretraining with multi-modal inputs.

Finally, MixMAE, which incorporates both Sentinel-2 and Sentinel-1 as inputs, allowing the model to address misalignment errors, achieved the highest finetuning score and outperformed all baselines overall. Effectively, this model learned two transformations: Sentinel-2 to Sentinel-1, and Sentinel-1 to Sentinel-2. Comparing to AuxMAE, which only learns to predict Sentinel-2 to Sentinel-1, the mixing model has the ability to learn a better representation of geospatial images, while reducing misalignment errors.

Table 2. Result summary for training the pretrained models on three downstream tasks: BigEarthNet (BEN) (classification), Sen1Floods (segmentation), and UrbanSRSeg8 (segmentation). We compare the capabilities of three models: (1) Baseline, which includes a single task of Sentinel-2 to Sentinel-2, (2) AuxMAE, which includes two tasks: Sentinel-2 to Sentinel-2 and Sentinel-2 to Sentinel-1, and (3) MixMAE, where the input is a random mixture of Sentinel-2 and Sentinel-1 patches, and the model reconstructs the two full images. We compare two scenarios: Finetune and Frozen, differing in whether the encoder is frozen during training of the downstream task or not.

Scenario	Model	BEN	Sen1Floods	UrbanSRSeg8
Finetune	Baseline	91.05	84.66	50.71
	AuxMAE	91.23	**85.26**	50.21
	MixMAE	**91.56**	**85.29**	**51.23**
Frozen	Baseline	76.05	82.28	44.80
	AuxMAE	**78.99**	**83.38**	**47.44**
	MixMAE	**79.11**	80.5	**47.36**

5. Conclusion

In this paper, we introduced two novel self-supervised learning approaches, AuxMAE and MixMAE, designed to integrate data from multiple modalities in the remote sensing domain, where both labeled and modality-aligned data are limited. The evaluation on three distinct datasets (BigEarthNet, Sen1Floods11, and UrbanSRSeg8) demonstrates that both AuxMAE and MixMAE outperform baseline models trained on single-source data, particularly in low-resource (frozen encoder) scenarios. Interestingly, MixMAE achieved the best overall performance across all datasets in the finetune scenario. Its ability to encode mixed-source information offers the potential for improved spatial understanding and robustness to misalignment.

The main direction for future exploration would be to explore additional modalities beyond Sentinel-1 and Sentinel-2 data, such as high-resolution RGB imagery or hyperspectral data. Multiple modalities could provide further insight into how different sources complement each other for diverse applications like flood detection, urban monitoring, and biodiversity assessment. For example, utilizing high resolution imagery could drive the models into an even better generalization capabilities. While in AuxMAE adding data sources could be done easily, it is not the case in MixMAE, where the mixing strategy between sources has a major effect, and should be further investigated.

References

[1] Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *Advances in Neural Information Processing Systems*, 36:79142–79154, 2023. 4, 5

[2] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*, 2020. 1

[3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 2

[4] Kumar Ayush, Burak Uzkent, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Efficient poverty mapping from high resolution remote sensing images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12–20, 2021. 1

[5] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2, 3

[6] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022. 2

[7] 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 (CVPR) Workshops*, June 2020. 1, 2, 6

[8] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021. 1

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 1

[10] Qiuyang Chen, Xenofon Karagiannis, and Simon M Mudd. Detecting floods from cloudy scenes: A fusion approach using sentinel-1 and sentinel-2 imagery. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1, 2

[13] 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, 2021. 1, 2

[14] Yuxing Chen, Maofan Zhao, and Lorenzo Bruzzone. A novel approach to incomplete multimodal learning for remote sensing data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[15] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1, 2, 3, 4, 6

[16] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[17] Fahimeh Fooladgar and Shohreh Kasaei. Multi-modal attention-based fusion model for semantic segmentation of rgb-depth images. *arXiv preprint arXiv:1912.11691*, 2019. 2

[18] Shubhika Garg, Ben Feinstein, Shahar Timnat, Vishal Batchu, Gideon Dror, Adi Gerzi Rosenthal, and Varun Gulshan. Cross-modal distillation for flood extent mapping. *Environmental Data Science*, 2:e37, 2023. 2

[19] Giorgio Giannone and Boris Chidlovskii. Learning common representation from rgb and depth images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2

[21] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024. 1, 2

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 4

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[24] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022. 5

[25] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Multimodal self-supervised representation learning for earth ob-

servation. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3241–3244. IEEE, 2021. 1, 2, 5

[26] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*, 2022. 1, 2, 7

[27] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 1

[28] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 1

[29] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 1

[30] Wenyuan Li, Hao Chen, and Zhenwei Shi. Semantic segmentation of remote sensing images with self-supervised multitask representation learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6438–6450, 2021. 2

[31] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 2

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

[33] Jorge Andres Chamorro Martinez, Laura Elena Cué La Rosa, Raul Queiroz Feitosa, Ieda Del'Arco Sanches, and Patrick Nigri Happ. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:188–201, 2021. 1

[34] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 1

[35] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. *arXiv preprint arXiv:2405.02771*, 2024. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[37] 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. 1, 2

[38] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 169:421–435, 2020. 1

[39] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 7

[40] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021. 2

[41] 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. 1, 2, 6, 7

[42] Aidan M Swope, Xander H Rudelis, and Kyle T Story. Representation learning for remote sensing: An unsupervised sensor fusion approach. *arXiv preprint arXiv:2108.05094*, 2021. 2

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 1, 2

[44] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023. 3

[45] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–5, 2018. 1

[46] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023. 1, 2

[47] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12275–12284, 2020. 1

[48] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 1