# MMEarth: Exploring Multi-Modal Pretext Tasks For Geospatial Representation Learning

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang

University of Copenhagen, Denmark
vishalned@gmail.com, nila@di.ku.dk

**Abstract.** The volume of unlabelled Earth observation (EO) data is huge, but many important applications lack labelled training data. However, EO data offers the unique opportunity to pair data from different modalities and sensors automatically based on geographic location and time, at virtually no human labor cost. We seize this opportunity to create *MMEarth*, a diverse multi-modal pretraining dataset at global scale. Using this new corpus of 1.2 million locations, we propose a *Multi-Pretext Masked Autoencoder (MP-MAE)* approach to learn general-purpose representations for optical satellite images. Our approach builds on the ConvNeXt V2 architecture, a fully convolutional masked autoencoder (MAE). Drawing upon a suite of multi-modal pretext tasks, we demonstrate that our MP-MAE approach outperforms both MAEs pretrained on ImageNet and MAEs pretrained on domain-specific satellite images. This is shown on several downstream tasks including image classification and semantic segmentation. We find that pretraining with multi-modal pretext tasks notably improves the linear probing performance compared to pretraining on optical satellite images only. This also leads to better label efficiency and parameter efficiency which are crucial aspects in global scale applications. [1]

**Keywords:** representation learning, self-supervised learning, multi-modal, multi-task, masked autoencoder, Earth observation, remote sensing, satellite images, Sentinel-2

## 1 Introduction

Learning representations through self-supervision has seen major advancements in recent years [2,5,22,32,34]. This gives hope that even applications with limited supervision data may profit from the promises of the deep learning revolution. However, some self-supervised techniques still fall short when evaluated on image domains other than ImageNet, *e.g.*, ground-level images from the natural

[1] The MMEarth dataset is available on the project page: vishalned.github.io/mmearth. The dataset construction code is available here: github.com/vishalned/MMEarth-data. The MP-MAE code for training and evaluation is available here: github.com/vishalned/MMEarth-train.

world [48]. Thus, to develop generalizing representation learning techniques, it is necessary to have suitable large-scale datasets from different domains. As an alternative to the "pure" self-supervision paradigm, multi-modal data offer great potential for learning good semantic representations [4, 31, 37]. Aligned multi-modal datasets are key for advancing two major research directions in computer vision: i) exploiting multi-modal data for representation learning, and ii) advancing representation learning to exploit multi-modal data for inference.

One domain that is particularly well-suited for studying the potential of multi-modal data is Earth Observation (EO). Geolocated data from different sensors can automatically be aligned at virtually no human labor cost. Aligning such EO data at large scale has become substantially easier thanks to cloud-based platforms like Google Earth Engine [21] and Microsoft Planetary Computer [30]. Furthermore, real-world remote sensing applications are often hampered by a lack of high-quality reference data. Human annotation of satellite data is not always possible from the images alone, for example when mapping carbon stocks or species abundance. Such applications require field measurements by experts that are costly, time consuming, and do not scale to large areas. Other high-quality measurement systems, such as airborne or spaceborne LiDAR, can provide a means to scale reference data, but still provide a biased sample of the Earth, due to the limits of global sampling. Here, we present a global dataset called *MMEarth* at the scale of ImageNet-1k [13] with multi-modal data for 1.2M locations. We also release our data collection framework for collecting data from different sensors given a list of geospatial coordinates. While MMEarth can be used to advance both aforementioned research directions, we demonstrate its potential for the first goal of exploiting multi-modal data for improving representation learning. Specifically, our goal is to learn general-purpose representations for optical satellite images from the Sentinel-2 mission that are predictive for a wide range of downstream tasks including crop type, landcover, and climate zone classification. Thus, in contrast to other work in this area, we neither aim to generalize to other sensor inputs such as Scale-MAE [38] nor to make use of multiple modalities during inference in downstream tasks such as MultiMAE [4] or 4M [31]. As shown by previous work, specializing a model to a particular input modality can be beneficial compared to models that aim to generalize to multiple input modalities [31].

To explore the potential of multiple pretext tasks, we build on fully convolutional masked autoencoders (FCMAE) and extend the ConvNeXt V2-MAE approach [53] with multi-modal reconstruction tasks for pretraining. As opposed to MAEs based on vision transformers, the FCMAE is well suited for downstream tasks that require different input image sizes, which is inevitable when evaluating the representations on existing EO benchmarks [27]. Intuitively, such a multi-task pretraining strategy with a shared encoder should lead to representations that generalize better to new downstream tasks that are not known at pretraining time. We evaluate the generalization potential using Sentinel-2 downstream tasks from GEO-Bench [27], including image-level multi-class and

multi-label classification as well as pixel-level semantic segmentation tasks. Our contributions can be summarized as follows:

1. **MMEarth dataset** - A global dataset for multi-modal and geospatial representation learning. It consists of 12 modalities including pixel-level and image-level modalities from 1.2 million locations.
2. **Method** - We explore the effect of multi-modal pretext tasks for both pixel-level and image-level modalities when learning general purpose representations for interpreting optical images from Sentinel-2. Therefore, we propose a Multi-Pretext Masked Autoencoder (MP-MAE) approach to using MMEarth to learn better representations for Sentinel-2 optical images.
3. **Results** - We show that pretraining MP-MAE on MMEarth outperforms not only MAEs pretrained on ImageNet but also on Sentinel-2 images. While both fine-tuning and linear probing performance are improved, linear probing benefits particularly from the multi-modal pretext tasks.

## 2 Related work

### 2.1 Masked image modelling

Masked Image Modelling (MIM) can see its roots in denoising autoencoders that are trained to reconstruct original images from partially corrupted inputs [50]. This idea was extended to pretrain CNN autoencoders by reconstructing a large area in an image with the aim of learning better representations by conditioning a portion of the image on its surroundings [35]. With the success of masked language modelling (for example, BERT [14]) and the rise of Vision Transformers (ViT) [16], Masked Autoencoder (MAE) approaches have been developed that reconstruct masked input patches [16, 22, 54]. When masking a large fraction of the input sequence, MIM is not only a successful pretext task to learn semantic representations, but also leads to a more efficient ViT encoder [22,54]. This masking strategy has also been extended to spatio-temporal modelling to learn from videos [6, 18, 45, 52] as well as to learn representations for 3D point clouds [56]. To combine this random masking strategy with the inductive bias of CNNs, we use ConvNeXt V2 [53], which is a Fully Convolutional MAE (FCMAE). The encoder is implemented with sparse convolutions [9] to improve efficiency like in the ViT-based MAE. ConvNeXt V2 also uses a learnable mask token which serves as the input to the convolutional decoder. Instead of using self-attention, the decoder is based on dense convolutions with kernels that cover the full spatial extent of the image. This architecture design pretrained on ImageNet yields a high-performing model family of different sizes, ranging from 3.7M ('Atto') to 650M parameters ('Huge'). We adapt the ConvNeXt V2 architecture design to satellite images and pretrain on our MMEarth dataset to demonstrate the potential of multi-modal EO data for representation learning.

### 2.2 Multi-modal representation learning

Multi-sensory information has been argued to help humans (and animals) to understand the natural world [41], which motivates the development of machine

learning approaches that learn representations from multi-modal data. Successful approaches such as CLIP [37] are based on a contrastive learning objective while learning separate encoders for each input modality. This approach has been extended to learn from images, text, and audio extracted from videos [1]. While contrastive approaches usually rely on aligned samples from various modalities, non-contrastive approaches were proposed to *e.g.* learn image-text representations with a single encoder to avoid the need for paired image-text data [19]. It has been shown that visual tasks are (more or less) related and that it can be more data-efficient to learn multiple tasks jointly [57]. In line with these observations, multi-task self-training [20] leverages pseudo-labels for multi-task learning to yield improved general purpose representations. In addition to such multi-task objectives, MIM has been extended to the multi-modal image setting [4, 31, 52]. Masked feature prediction [52] has been demonstrated as an effective pretext task, *e.g.* by reconstructing the Histograms of Oriented Gradients (HOG) instead of the original image, arguing that the local contrast normalization plays a crucial role. MultiMAE [4] extends the MAE approach [22] to multiple input and output modalities using RGB, semantic label masks, and depth. Recently, 4M [31] proposed a masked modelling framework to train a single autoencoder for a range of input and output modalities. Instead of reconstructing each modality in the original space (*e.g.* pixels for RGB), each modality is first mapped to discrete tokens using a modality-specific autoencoder. These tokens are then used as the input and output to a transformer-based autoencoder trained with masked modelling, leading to impressive any-to-any reconstruction results. However, their ablation study demonstrates that pretraining with a single input modality can outperform a model trained with multiple input modalities. While previous works have focused on image modalities that provide pixel-level data, our approach makes use of both pixel-level and image-level modalities.

### 2.3 Representation learning in Earth observation

**Supervised learning at large scale.** In applications where labelled training data is abundant, supervised learning has been used in a range of applications like tree segmentation [47], marine debris detection [40], and snow depth estimation [12]. However, only a few studies have scaled supervised deep learning to global scales to map land cover [8], canopy height [28], or solar panels [26]. Additionally, large annotated EO datasets have been created to replace ImageNet with a domain-specific pretraining dataset [7,10,42,43,51]. Some bi-modal datasets provide Sentinel-2 optical and Sentinel-1 SAR images in Europe [43] and globally around cities [51]. The benefit of multi-modal input data has been studied for landcover mapping at the scale of Europe in a fully supervised setting [33]. In contrast, MMEarth provides a global pretraining dataset with 12 modalities balanced across 14 biomes [15].

**Self-supervised learning.** Self-supervised approaches developed on ground-level computer vision benchmarks (like ImageNet) can be applied to satellite

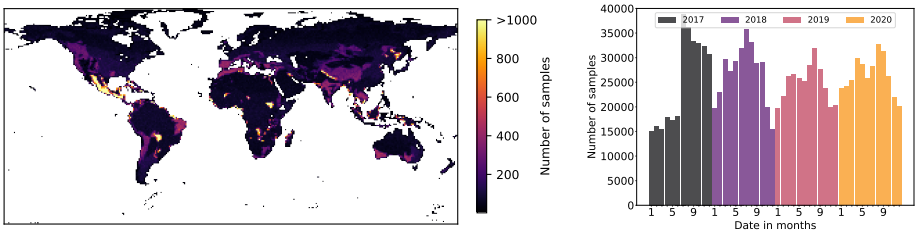
images [44, 51]. However, these approaches might be suboptimal due to the differences in ground-level and satellite images. Therefore, the EO community has started to explore the unique opportunities offered by EO data. We consider three major aspects of EO data for self-supervised representation learning: *geolocation*, *time*, and *modality*.

**Geolocation.** In an approach called geography-aware self-supervised learning (GASSL) [3], the contrastive learning objective leverages the spatio-temporal structure of EO data. Using spatially aligned images, positive image pairs can be created by using observations at different times, avoiding the need for artificial augmentations. In addition, localization is used as a pretext task by predicting the clusters of sampled locations. SatCLIP [25] focuses on learning location embeddings in a contrastive manner by learning a Sentinel-2 encoder and a location encoder with a contrastive objective. **Time.** Seasonal contrast (SeCo [29]) is another temporal contrastive approach that learns multiple embedding subspaces that are invariant to either seasonal changes, synthetic augmentations, or all transformations. Alternatively, image time series can be exploited in masked modelling as was demonstrated in SatMAE [11]. Presto [46] is a masked modelling approach that focuses on learning representations for multi-modal time series using a light-weight transformer architecture that only relies on pixel time series inputs and ignores textural features. **Modality.** Scale-MAE [38] aims to generalize to different input resolutions by conditioning on the ground sampling distance (GSD, size of a pixel on the ground). Most related to our work is SatlasPretrain [7], a large multi-task dataset for optical satellite images including Sentinel-2. Their goal was *"to label everything that is visible in a satellite image"* using various data sources (Open Street Map, lidar scans, landcover maps) and human annotators (domain experts and Mechanical Turk workers). Our MMEarth dataset creation differs, as it is fully automatic without relying on human annotators and also considers modalities that are not *explicitly visible to humans* in optical images. Such modalities include everything that cannot be annotated by a human like radar, temperature, precipitation, or canopy height. We hypothesize that predicting such cross-modal relations from images requires the extraction of semantic features.

All these works provide important insights into how to make better use of satellite data and its metadata in representation learning. Our work extends along the last aspect — exploiting multiple modalities — whereas space and time serve as the underlying structure for pairing data from different sources automatically without relying on human annotation. Furthermore, we treat geolocation and time as two additional modalities in our framework.

## 3 MMEarth dataset

MMEarth contains data for 1.2 million locations distributed around the world, making its optical image count comparable to that of ImageNet-1K [13]. At each location, data from 12 geo-aligned modalities were collected, grouped into pixel-level and image-level modalities as per Table 1. The six pixel-level modalities




**(a)** Spatial distribution. **(b)** Temporal distribution.

**Fig. 1: MMEarth dataset coverage.** With a balanced sampling scheme across 14 biomes and 4 years, we collected aligned multi-modal data from 12 modalities using the Google Earth Engine platform [21] at 1.2M locations.

**Table 1: Modalities in the MMEarth dataset.** Each location provides 12 aligned modalities with a total of 46 bands. The upper half of the modalities provide dense data at the pixel level and the bottom half at the image level.

| | Sensor/Product | Description | Data type | #Bands | Band information |
|---|---|---|---|---|---|
| Pixel-level | Sentinel-2 | Optical | continuous | 12 | multi-spectral B1-B12 for L1C/L2A product |
| | Sentinel-1 | SAR | continuous | 8 | VV, VH, HV, HH for ascending/descending orbit |
| | Aster DEM | Elevation | continuous | 2 | elevation, slope |
| | ETH-GCHM | Vegetation height | continuous | 2 | canopy height, uncertainty (STD) |
| | Dynamic World | Landcover | categorical | 1 | 9 landcover categories |
| | ESA World Cover | Landcover | categorical | 1 | 11 landcover categories |
| Image-level | Biome | Landcover | categorical | 1 | 14 terrestrial ecosystem categories |
| | Ecoregion | Landcover | categorical | 1 | 846 ecoregion categories |
| | ERA5 temperature | Climate reanalysis | continuous | 9 | mean, min, and max of [year, month, previous month] |
| | ERA5 precipitation | Climate reanalysis | continuous | 3 | total precipitation of [year, month, previous month] |
| | Geolocation | Latitude, Longitude | continuous | 4 | cyclic encoding of latitude and longitude |
| | Date (Sentinel-2) | Month of the year | continuous | 2 | cyclic encoding of the month |

represent raster data of size $128 \times 128$ pixels which capture $1.28 \times 1.28$ km on the ground (*e.g.*, Sentinel-2, Sentinel-1, Aster DEM, Dynamic World, and ESA World Cover). The remaining six image-level modalities represent scalar values for each location (*e.g.*, Biome, Ecoregion, ERA5 temperature, ERA5 precipitation, Geolocation, and Sentinel-2 observation date). The size of the input images is large enough for typical downstream EO tasks [27] and yields a reasonable dataset volume. MMEarth was sampled uniformly across 14 biomes (Fig. 1, *e.g.*, Mangroves, Temperate Conifer Forests, Tundra, etc.) [15]. To increase diversity, we considered data from the four years 2017–2020 (Fig. 1). Further, we ensured that time-critical modalities were collected around the Sentinel-2 observation date, which serves as the reference. Additionally, all pixel-level modalities were re-projected to the Sentinel-2 10m grid, if needed, yielding a harmonized data cube. In addition to the full dataset, we provide two subsets to facilitate research in multi-modal representation learning with limited compute resources: *MMEarth100k* (100k locations at $128 \times 128$ pixels) and *MMEarth64* (1.2M locations center-cropped to $64 \times 64$ pixels); see supplementary material for details. All the data were downloaded from Google Earth Engine (GEE) [21]. In the following subsections, we provide a brief overview of the pixel-level and image-

level modalities listed in Table 1. More details are provided in the supplementary material.

### 3.1 Pixel-level modalities

**Sentinel-2.** An optical satellite mission that provides global coverage at least every 5 days at 10m resolution. Its multi-spectral sensor comprises of 12 bands, with four bands RGBN (RGB and near infrared) at 10m and near and short-wave infrared bands at 20m and 60m resolution, respectively. We upsample all bands to the 10m grid. Sentinel-2 provides two processing levels: top of atmosphere reflectance (L1C) and bottom of atmosphere reflectance (L2A). L2A is atmospherically corrected, harmonizing the spectral data, but is not yet available for all years on a global scale. To ensure global coverage for the years 2017–2020, MMEarth includes both L2A and L1C data. **Sentinel-1.** Synthetic Aperture Radar (SAR) data at 10m resolution, collected in two viewing directions along an ascending and descending orbit, resulting in four bands each. **Aster GDEM.** A Global Digital Elevation Model derived from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). The Aster GDEM data covers 99% of the Earth's landmass at 15m resolution. We include elevation and slope. **ETH-GCHM.** A 10m Global Canopy Top Height Map [28] for the year 2020 derived from Sentinel-2 and sparse vertical structure data from GEDI [17], a spaceborne LiDAR mission. **Dynamic World.** A land cover dataset that contains semantic segmentation of Sentinel-2 images for nine categories at 10m resolution. We aggregate the per-image maps for each year to reduce noise in the data. **ESA World Cover.** A 10m global land cover map based on Sentinel-1 and Sentinel-2 data with 11 categories for the year 2020.

### 3.2 Image-level modalities

**Biome.** A biome is a geographical region with specific vegetation, climate and animal life. We use the RESOLVE Ecoregions dataset [15] that consists of 14 terrestrial biomes. **Ecoregion.** The RESOLVE ecoregions [15] include 846 terrestrial ecoregions. These are a more fine-grained classification than biomes. For example, the ecoregions *Central African Mangroves* and *Indochina Mangroves* belong to the biome *Mangroves*. **ERA5 Temperature and Precipitation.** ERA5 reanalysis data provides climate information back to 1950. Based on the Sentinel-2 observation date, we collect the corresponding mean, min, and max temperature as well as the total precipitation for the month, previous month, and year. **Geolocation.** We use cyclic encodings for the latitude and longitude of each image center. This ensures that for the longitude the cyclic property is maintained (transition from West to East). **Date.** We use a cyclic encoding of the month of the Sentinel-2 observation date as a proxy for the season.

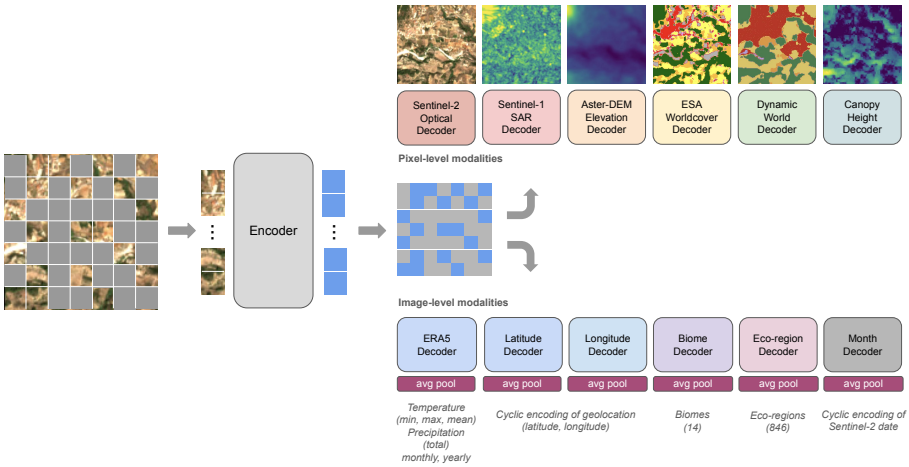



**Fig. 2: Multi-Pretext Masked Autoencoder (MP-MAE).** Our approach extends Masked Autoencoders, which reconstruct only the input image, by incorporating multiple pretext tasks using aligned pixel-level as well as image-level modalities.

## 4 Methodology

### 4.1 Multi-Pretext Masked Autoencoder

In this section, we describe the proposed Multi-Pretext Masked Autoencoder (MP-MAE) approach as illustrated in Fig. 2. It builds on the promising results of MIM with the ConvNeXt V2 architecture.

**Encoder.** The input image $x$ is divided into patches (or tokens) on which a random mask is applied to yield a masked image $x'$. The encoder $f_\theta$ takes the visible patches as the input and returns an encoding $z$ for every visible patch. Here, the ConvNeXt V2 [53] architecture serves as the backbone which is implemented with sparse convolutions [9] to improve efficiency. We investigated two key modifications to adapt the ConvNeXt V2 architecture to medium-resolution satellite images. First, we adapted the patch size. The original ConvNeXt V2-MAE has been developed for pretraining on ImageNet and assumes input images of size $224 \times 224$ pixels, which are split into $7 \times 7$ patches of size $32 \times 32$ pixels. MMEarth images are of size $128 \times 128$ pixels, and we randomly crop them to $112 \times 112$ pixels when feeding them into the MP-MAE encoder, which has a reduced patch size of $16 \times 16$ pixels to preserve the $7 \times 7$ patch layout. This adjustment is crucial as the number of patches is coupled with the optimal masking ratio. Second, we avoid early downsampling by modifying the first layer. In the original ConvNeXt V2 encoder, the first layer is a learned convolutional downsampling layer, which yields feature maps at $1/4$ of the input resolution. While this approach might be efficient and loses only little spatial information that is not needed for

ImageNet classification, it may not be optimal for medium-resolution satellite images such as Sentinel-2 and for pixel-level downstream tasks such as semantic segmentation with U-Net architectures [39], where high-resolution features can inform same-resolution representations via the skip connections. Therefore, we replace the first layer with an initial convolutional layer with kernel size 3 and stride 1 to learn feature maps at the input resolution followed by a depth-wise convolutional downsampling layer.

**Decoders.** We follow the ConvNeXt V2-MAE [53] decoder design and use shallow decoders with one block. We apply the same random mask to all pixel-level modalities and learn a separate decoder $h_t$ for each pretext task. We consider $T$ pretext tasks and treat most modalities as individual tasks, but group the climate variables and split latitude and longitude (see Fig. 2). To reconstruct the masked patches, a learnable mask token is used as a placeholder when combining the embedding tokens to get a dense 2D-input for the decoders. Therefore, the decoders are based on standard convolutions. For the image-level modalities, we introduce a global average pooling over the decoded non-visible patches before the final linear prediction layer. Hence, we follow the same strategy as for the pixel-level modalities, wherein the encoder is encouraged to learn representations that provide the semantic context that is needed to decode the masked tokens.

### 4.2 Multi-task loss

All pretext task targets $y_t$ are reconstructed by applying the same random mask to prevent the model from learning any shortcuts between the input and the targets. We follow prior work and compute the losses only over the non-visible patches. We experimented with two settings for the multi-task loss function [24, 49]

$$\mathcal{L} = \sum_{t=1}^{T} \frac{1}{\hat{\sigma}_t^2} \mathcal{L}_t \left( h_t(f_\theta(x)), y_t \right) + \log \hat{\sigma}_t, \tag{1}$$

with task-specific loss $\mathcal{L}_t$ (*e.g.*, cross entropy for classification and mean squared error for regression) and variance $\hat{\sigma}_t^2$. First, we applied equal weighting with $\hat{\sigma}_t^2 = 1$ for all decoders, followed by task-uncertainty weighting where $\hat{\sigma}_t^2$ is learnable for each decoder [4,24]. The second setting was expected to aid in handling noisy targets, as task-specific weights naturally decrease for such pretext tasks [24].

## 5 Experimental setup

In our problem setup, the downstream tasks are not known at pretraining time. Given a pretrained encoder, we follow standard practices and evaluate fine-tuning (FT) and linear probing (LP) performance on new downstream tasks. While both approaches learn the last linear layer of a neural network from scratch, FT updates the pretrained encoder and LP uses a frozen encoder. Implementation details are provided in the supplementary material.

**Table 2: Downstream task datasets.** We evaluated on five different datasets which involve image-level classification (multi-label, multi-class) and pixel-level semantic segmentation. These datasets are modified versions from the GEO-Bench benchmark [27].

| Dataset Name | Description | Task | Categories | Input [px] | Train/Val/Test |
|---|---|---|---|---|---|
| BigEarth20k [42] | Landcover classification in Europe | Multi-label | 43 | 120×120 | 20k/1k/1k |
| So2Sat20k [58] | Global local climate zones classification | Multi-class | 17 | 32×32 | 20k/1k/1k |
| EuroSat2k [23] | Landcover classification in Europe | Multi-class | 10 | 64×64 | 2k/1k/1k |
| Cashew1k [55] | Cashew plantations in Benin (Africa) | Segmentation | 7 | 256×256 | 1.3k/400/50 |
| SAcrop3k [36] | Crop-type segmentation in South Africa | Segmentation | 10 | 256×256 | 3k/1k/1k |

### 5.1 Evaluation on downstream tasks

For downstream tasks, we considered all five Sentinel-2 datasets provided in the GEO-Bench [27] EO dataset collection[2]. These datasets provide a diverse set of tasks and geographical coverage (Global, Europe, Africa). GEO-Bench provides harmonized subsets of the original datasets with comparable training set sizes and balanced class distributions. For clarity, we refer to the datasets with the number of training samples to avoid confusion with prior work using the original datasets. We summarize the characteristics of these five datasets in Table 2 and provide more details in the following sections.

**Image classification: BigEarthNet** [42] is a multi-label land cover classification problem with 43 categories and data from 10 countries in Europe. While the original dataset contains a total of 590k Sentinel-2 image samples, the GEO-Bench version, referred to as 'BigEarth20k', contains 20k training samples. **So2Sat** [58] is a multi-class land cover classification task with 17 categories that classifies local climate zones. The data is collected globally from urban areas and the original data contains 400k pairs of Sentinel-2 and Sentinel-1 image samples. The GEO-Bench version contains 20k training samples and is referred to as 'So2Sat20k'. **EuroSat** [23] is a multi-class land use and land cover classification dataset covering Europe with 10 categories. The original dataset contains 27k image samples, and the GEO-Bench version 'EuroSat2k' has roughly 2k training samples.

**Semantic segmentation: Cashew Plantation** [55] is a semantic segmentation task with 7 categories that maps cashew plantations in Benin, Africa. The 'Cashew1k' consists of 1k training samples. **SA-Crop-Type** [36] is another semantic segmentation problem with 10 categories that maps crop type in South Africa. The 'SAcrop3k' consists of 3k training samples.

**U-Net evaluation for semantic segmentation:** We adopt a U-Net [39] architecture to evaluate the pretrained encoders. These U-Nets are fine-tuned in two phases, referred to as FT in the results. First, we trained the randomly initialized U-Net decoder with the frozen pretrained encoder for 50 epochs, and then fine-tuned the full model for another 150 epochs.

**Evaluation metrics:** Following [27], we use the micro-averaged overall accuracy (Acc.) for multi-class, the micro-averaged F1-score (F1) for multi-label classifi-

[2] We use the latest GEO-Bench version v1.0 in which datasets were class-balanced.

cation, and the macro-averaged intersection over union (IoU) for semantic segmentation tasks.

## 6 Results

We investigated two research questions: *How can domain specific pretraining improve representations?* and *How do multi-modal pretext tasks impact representations?*

### 6.1 How can domain specific pretraining improve representations?

**Switching to domain-specific data is not enough.** The results from Table 3 indicate that just replacing the ImageNet pretraining data with domain-specific optical satellite images does not improve downstream performance. We propose slightly adjusting the encoder design and the patch size to adapt to medium resolution satellite images like Sentinel-2 (10m pixels on the ground). Reducing the patch size from 32 to 16 or 8 pixels, for image sizes of 112 or 56 pixels, respectively, is crucial to adapt to the smaller image size of the medium resolution satellite images. This preserves the number of patches to be $7\times7$ as in the original ConvNeXt V2 ImageNet setting and is related to an optimal masking ratio [54]. With these modifications, domain-specific pretraining improved fine-tuning and linear probing performance across all datasets compared to pretraining on ImageNet. The improvement is smallest for FT on Cashew1k with 3pp and largest for FT on So2Sat with 12pp or 17pp, when training on image sizes of 112 and 56, respectively.

**Multi-spectral images are beneficial.** Since Sentinel-2 images provide 12 bands in total, we studied the impact of pretraining on multi-spectral images and found that this can improve downstream performance depending on the tasks (see Table 3). So2Sat20k is an exception where the model using only the RGB channels achieved high performance. The following multi-pretext results are based on using all Sentinel-2 bands as input.

### 6.2 How do multi-modal pretext tasks impact representations?

**Multi-modal pretext tasks improve fine-tuning and linear probing.** Using multi-pretext tasks improved both the FT and LP performance for all downstream tasks (see Table 4). While MAE are known for good FT performance, the gap with the LP performance remains large [22,53]. The representations learned from our multi-pretext objective notably improved LP performance and provide a mechanism towards closing this performance gap. We observed an improvement of ≈5pp on BigEarth20k and 8pp on So2Sat20k with respect to pretraining on Sentinel-2 only. Here, we report results obtained with the task uncertainty loss. We found that both loss settings work equally well on these downstream tasks (see supplementary). However, the learned task uncertainties provide interesting

**Table 3: Domain specific pretraining on optical images.** Fine-tuning (FT) and linear probing (LP) performance using MAE pretrained on ImageNet (RGB), Sentinel-2 RGB (MMEarth64-S2rgb, MMEarth-S2rgb), and Sentinel-2 multi-spectral (MMEarth64-S2). Lowering the patch size is crucial to adapt to the smaller medium resolution satellite images. Using all 12 bands improves accuracy. Domain-specific pretraining improves both FT and LP results.

| Pretrain data | Image size | Patch size | BigEarth20k (F1↑) FT/LP | So2Sat20k (Acc.↑) FT/LP | Cashew1k (IoU↑) FT | SAcrop3k (IoU↑) FT |
|---|---|---|---|---|---|---|
| ImageNet | 224 | 32 | 55.7/25.9 | 36.6/24.0 | 77.1 | 26.7 |
| MMEarth-S2rgb | 128 | 32 | 55.5/29.8 | 38.1/22.7 | 77.2 | 23.8 |
| MMEarth-S2rgb | 112 | 16 | 62.3/33.9 | 48.8/29.8 | 79.3 | 28.7 |
| MMEarth64-S2rgb | 56 | 8 | 64.9/34.4 | **54.1**/32.5 | **80.0** | 33.1 |
| MMEarth64-S2 | 56 | 8 | **65.1**/**36.2** | 48.7/**33.5** | 79.9 | **36.0** |

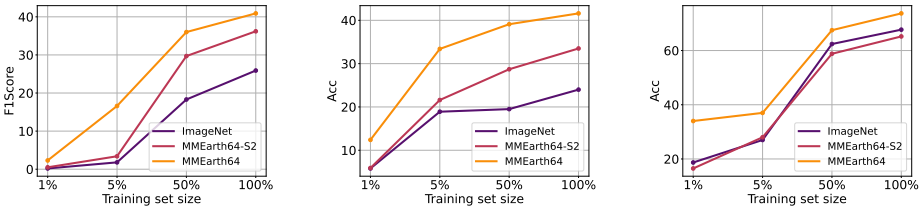
insights, *e.g.* the uncertainty for the latitude (N-S) being lower than for the longitude (E-W), which could be explained by the rainfall and vegetation gradients along the North-South axis shaping the appearance of landscapes.

**Pixel-level and image-level pretext tasks are useful.** To study the role of the two groups of pretext tasks, we ran experiments where the Sentinel-2 reconstruction task was extended with only one group (see Table 4). For LP we found that combining the tasks from both groups leads to the best performance on both datasets. For BigEarth20k using only one group did not improve LP performance over pretraining on Sentinel-2 only. For So2Sat20k image-level pretext tasks lead to slightly better LP performance, but pixel-level pretext tasks lead to better FT performance. While we had expected that pixel-level pretext tasks might be more beneficial for segmentation tasks, this was not confirmed. From these experiments we hypothesize that the semantics of the pretext tasks are more important than whether the tasks are pixel- or image-level.

**Multi-modal pretext tasks improve label efficiency.** Real-world applications often lack large amounts of labelled data. To explore the benefits of pretraining in a limited label regime, we evaluated the LP performance using stratified subsets of the fine-tuning datasets (GEO-Bench partitions). In the most extreme few-shot setting for BigEarth20k, there were only 200 training samples, *i.e.* 4–5 samples/category. For EuroSat2k the extreme few-shot setting has only 20 training samples with 2 samples/category. Results in Fig. 3 show that models pretrained with multi-modal pretext tasks on MMEarth outperformed models trained on only Sentinel-2 or only ImageNet in few-shot linear probing. While EuroSat is an easy task for supervised learning, GEO-Bench provides a rather small subset, which makes it a challenging few-shot task. For EuroSat2k, we found that general image features learned on ImageNet are a strong baseline working on par with our model trained on Sentinel-2 images only. Even in the extreme case when only using 1% of EuroSat2k, our multi-modal pretext tasks improve LP performance from 16.5% to 34% accuracy.

**Table 4: Multi-modal pretext tasks.** Fine-tuning (FT) and linear probing (LP) performance for multi-pretext pretraining (MMEarth64, MMEarth) and masked image pretraining (ImageNet, MMEarth64-S2). We also present multi-pretext results with subsets of modalities using only pixel-level (MMEarth64-PixelM) or only image-level (MMEarth64-ImageM) modalities in addition to the Sentinel-2 optical image reconstruction. Multi-pretext pretraining improves both FT and LP results.

| Pretrain | BigEarth20k (F1↑) FT/LP | So2Sat20k (Acc.↑) FT/LP | Cashew1k (IoU↑) FT | SAcrop3k (IoU↑) FT |
|---|---|---|---|---|
| ImageNet | 55.7/25.9 | 36.6/24.0 | 77.1 | 26.7 |
| MMEarth64-S2 | 65.1/36.2 | 48.7/33.5 | 79.9 | 36.0 |
| MMEarth64-PixelM | 67.5/36.6 | **58.2**/38.5 | **81.9** | 39.7 |
| MMEarth64-ImageM | **68.4**/36.6 | 56.3/39.2 | 81.4 | **39.8** |
| MMEarth64 | 68.2/40.9 | 54.0/41.7 | 81.6 | 39.7 |
| MMEarth | 67.1/**43.3** | 54.6/**43.8** | 79.8 | 38.2 |




**(a)** BigEarth20k **(b)** So2Sat20k **(c)** EuroSat2k

**Fig. 3: Label efficiency for few-shot downstream performance.** Linear probing performance for varying downstream dataset sizes. MP-MAE ('Atto') pretrained on ImageNet, MMEarth64-S2 (multi-spectral only), MMEarth64 (all multi-modal pretext tasks).

**Compares favorably to prior work.** We compare our MP-MAE method to a small, but most closely related, subset of prior work in Table 5. While all models rely on different pretraining datasets and pretraining strategies, we show that our model compares favorably to prior work on most tasks. The encoder in our MP-MAE consists of 3.7M parameters, which is much smaller compared to e.g. the 25M for ResNet-50. This supports our motivation for creating better domain-specific pretraining datasets and pretraining strategies.

### 6.3 Limitations

While this study introduces a new dataset and methodology to advance the use of multiple modalities for representation learning, some limitations exist. First, we rely on GEO-Bench for downstream evaluation with five datasets covering a range of tasks and geographic regions. Further evaluation on diverse application domains from various geographical regions could reveal insights about which

**Table 5: Comparison to prior work.** We compare our multi-pretext pretraining with prior work that makes use of different pretraining datasets and strategies. Our pretraining strategy achieves similar if not better performance compared to other methods that use a ResNet-50 backbone. We highlight the **best** and second best results.

| Method | Pretrain data (Num. locations) | Encoder (Num. parameters) | BigEarth20k (F1↑) FT/LP | So2Sat20k (Acc.↑) FT/LP | Cashew1k (IoU↑) FT | SAcrop3k (IoU↑) FT |
|---|---|---|---|---|---|---|
| GASSL [3] | fMoW (0.36M) | ResNet-50 (25M) | 53.6/34.1 | 42.4/35.4 | 79.0 | 21.1 |
| SeCo [29] | SeCo (0.20M) | ResNet-50 (25M) | 59.6/**44.5** | 45.2/39.9 | 77.5 | 22.7 |
| SatlasNet [7] | SatlasPretrain (0.86M) | ResNet-50 (25M) | 55.5/40.3 | 37.1/22.1 | 76.4 | 20.5 |
| MP-MAE (ours) | MMEarth (1.24M) | ConvNeXt V2-Atto (3.7M) | **67.1**/43.3 | **54.6**/**43.8** | **79.8** | **38.2** |

scenarios our proposed multi-modal pretext tasks are beneficial in at a global scale. Second, we run our pretraining for 200 epochs due to restricted compute resources. For the same reason we only train encoders with a limited number of parameters. Prior work exploring MAEs for representation learning pretrains for 800 or 1600 epochs. Third, our MP-MAE is not explicitly trained to be robust to varying numbers of input channels. We have trained the MP-MAE with multi-spectral Sentinel-2 images (12 bands). It needs to be investigated how our models transfer to downstream tasks that use e.g. RGB only. While this seems to be an artefact of benchmark datasets, rather than a real-world scenario when working with Sentinel-2 images, prior work proposed input masking strategies to make a model robust against missing channels [46], which could be built into our framework. Fourth, we are not making use of the validation dataset for early stopping during the FT/LP evaluation, which could improve results overall. Lastly, we only reported single trials. Running the FT/LP experiments multiple times would allow us to average out random fluctuations.

## 7 Conclusion

Prior work has mostly focused on exploiting *geolocation* and *time* for representation learning with EO data. We move a step forward and present MMEarth, a global *multi-modal* pretraining dataset for geospatial representation learning. With 12 modalities sampled from 1.2M locations, we propose a Multi-Pretext Masked Autoencoder (MP-MAE) approach based on ConvNeXt V2 to study the potential of aligned multi-modal data to learn better representations for Sentinel-2 optical satellite images. First, we show how to adapt the ConvNeXt V2 encoder to improve domain-specific pretraining with medium-resolution satellite images. Second, we demonstrate that multi-modal pretext tasks are beneficial for learning better representations for Sentinel-2. Both fine-tuning and linear probing are improved; the latter profits substantially in the few-shot setting. This is promising for real-world applications with limited training data. However, further research is needed to close the gap between the fine-tuning and linear probing performance.

**Acknowledgments.** We thank Lucia Gordon for the valuable feedback. We greatly appreciate the open data policies of the Copernicus program and its partners ESA and ECMWF. We thank Google Earth Engine for hosting the data and providing free access. This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1. The authors AK, CI, and NL acknowledge support by the research grant DeReEco (grant number 34306) from Villum Foundation. SO and CI acknowledge support by the research grant Global Wetland Center (grant number NNF23OC0081089) from Novo Nordisk Foundation. CI and SB acknowledge support by the European Union project ELIAS (grant agreement number 101120237). We thank the Danish e-Infrastructure Consortium (DeiC), Martin Brandt, and Konrad Schindler for their support with computing resources.

## References

1. Argaw, D.M., Lee, J.Y., Woodson, M., Kweon, I.S., Caba Heilbron, F.: Long-range multimodal pretraining for movie understanding. In: International Conference on Computer Vision (ICCV). IEEE (2023)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Computer Vision and Pattern Recognition (CVPR). pp. 15619–15629 (2023)
3. Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: International Conference on Computer Vision (ICCV). pp. 10181–10190 (2021)
4. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision (ECCV). pp. 348–367. Springer (2022)
5. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210 (2023)
6. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint (2024)
7. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: Satlaspretrain: A large-scale dataset for remote sensing image understanding. In: International Conference on Computer Vision (ICCV). pp. 16772–16782 (2023)
8. Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., et al.: Dynamic world, near real-time global 10 m land use land cover mapping. Scientific Data **9**(1), 251 (2022)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR). pp. 3075–3084 (2019)
10. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Computer Vision and Pattern Recognition (CVPR). pp. 6172–6180 (2018)
11. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral

satellite imagery. Advances in Neural Information Processing Systems (NeurIPS) **35**, 197–211 (2022)
12. Daudt, R.C., Wulf, H., Hafner, E.D., Bühler, Y., Schindler, K., Wegner, J.D.: Snow depth estimation at country-scale with high spatial and temporal resolution. ISPRS Journal of Photogrammetry and Remote Sensing **197**, 105–121 (2023)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Ieee (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
15. Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.P.B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F., Saleem, M.: An ecoregion-based approach to protecting half the terrestrial realm. BioScience **67**(6), 534–545 (2017)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
17. Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., et al.: The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. Science of remote sensing **1**, 100002 (2020)
18. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (NeurIPS) (2022)
19. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: Multimodal masked autoencoders learn transferable representations (2022)
20. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. In: International Conference on Computer Vision (ICCV). pp. 8856–8865 (2021)
21. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R.: Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment **202**, 18–27 (2017)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
23. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)

24. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Computer Vision and Pattern Recognition (CVPR). pp. 7482–7491 (2018)
25. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M.: Satclip: Global, general-purpose location embeddings with satellite imagery. arXiv preprint arXiv:2311.17179 (2023)
26. Kruitwagen, L., Story, K., Friedrich, J., Byers, L., Skillman, S., Hepburn, C.: A global inventory of photovoltaic solar energy generating units. Nature **598**(7882), 604–610 (2021)
27. Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E.D., Kerner, H., Lütjens, B., Irvin, J.A., Dao, D., Alemohammad, H., Drouin, A., Gunturkun, M., Huang, G., Vazquez, D., Newman, D., Bengio, Y., Ermon, S., Zhu, X.X.: GEO-Bench: Toward foundation models for earth monitoring. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2023)
28. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the earth. Nature Ecology & Evolution **7**(11), 1778–1789 (2023)
29. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: International Conference on Computer Vision (ICCV). pp. 9414–9423 (2021)
30. Microsoft Open Source, McFarland, M., Emanuele, R., Morris, D., Augspurger, T.: microsoft/planetarycomputer: October 2022 (Oct 2022), https://doi.org/10.5281/zenodo.7261897
31. Mizrahi, D., Bachmann, R., Kar, O.F., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4M: Massively multimodal masked modeling. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
32. Mohamed, A., Lee, H., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.W., Livescu, K., Maaløe, L., Sainath, T.N., Watanabe, S.: Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1179–1210 (2022)
33. Mommert, M., Kesseli, N., Hanna, J., Scheibenreif, L., Borth, D., Demir, B.: Benge: Extending bigearthnet with geographical and environmental data. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 1016–1019. IEEE (2023)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
35. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Computer Vision and Pattern Recognition (CVPR). pp. 2536–2544 (2016)
36. Planet, Radiant Earth Foundation, Western Cape Department of Agriculture, German Aerospace Center (DLR): A fusion dataset for crop type classification in Western Cape, South Africa (2021). https://doi.org/10.34911/RDNT.GQY868
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machien Learning (ICML). pp. 8748–8763. PMLR (2021)
38. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked au-

toencoder for multiscale geospatial representation learning. In: International Conference on Computer Vision (ICCV). pp. 4088–4099 (2023)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241. Springer (2015)
40. Rußwurm, M., Venkatesa, S.J., Tuia, D.: Large-scale detection of marine debris in coastal areas with sentinel-2. Iscience **26**(12) (2023)
41. de Sa, V.R., Ballard, D.H.: Category learning through multimodality sensing. Neural Computation **10**(5), 1097–1117 (1998)
42. Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 5901–5904. IEEE (2019)
43. Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V.: Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **9**(3), 174–180 (2021)
44. Tolan, J., Yang, H.I., Nosarzewski, B., Couairon, G., Vo, H.V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., Bojanowski, P., Johns, T., White, B., Tiecke, T., Couprie, C.: Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. Remote Sensing of Environment **300**, 113888 (2024)
45. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (NeurIPS) (2022)
46. Tseng, G., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H.: Lightweight, pre-trained transformers for remote sensing timeseries. arXiv preprint arXiv:2304.14065 (2023)
47. Tucker, C., Brandt, M., Hiernaux, P., Kariryaa, A., Rasmussen, K., Small, J., Igel, C., Reiner, F., Melocik, K., Meyer, J., et al.: Sub-continental-scale carbon stocks of individual trees in african drylands. Nature **615**(7950), 80–86 (2023)
48. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: Computer Vision and Pattern Recognition (CVPR). pp. 12884–12893 (2021)
49. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3614–3633 (2021)
50. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: International Conference on Machien Learning (ICML). pp. 1096–1103. ACM (2008)
51. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. arXiv preprint arXiv:2211.07044 (2022)
52. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Computer Vision and Pattern Recognition (CVPR). pp. 14668–14678 (2022)
53. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In: Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142 (2023)

54. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMIM: A simple framework for masked image modeling. In: Computer Vision and Pattern Recognition (CVPR). pp. 9653–9663 (2022)
55. Yin, L., Ghosh, R., Lin, C., Hale, D., Weigl, C., Obarowski, J., Zhou, J., Till, J., Jia, X., You, N., Mao, T., Kumar, V., Jin, Z.: Mapping smallholder cashew plantations to inform sustainable tree crop expansion in benin. Remote Sensing of Environment **295**, 113695 (2023)
56. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In: Computer Vision and Pattern Recognition (CVPR). pp. 19313–19322 (2022)
57. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Computer Vision and Pattern Recognition (CVPR). pp. 3712–3722 (2018)
58. Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., Wang, Y.: So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **8**(3), 76–89 (2020)