EarthView: A Large Scale Remote Sensing Dataset for Self-Supervision

Diego Velazquez[1,4], Pau Rodriguez López[1,5], Sergio Alonso[2], Josep M. Gonfaus[2], Jordi Gonzalez[1], Gerardo Richarte[2], Javier Marin[2], Yoshua Bengio[3], Alexandre Lacoste[4]
[1]Computer Vision Center [2]Satellogic [3]Mila, Université de Montréal [4]ServiceNow Research [5]Apple Research

Abstract

This paper presents EarthView, a comprehensive dataset specifically designed for self-supervision on remote sensing data, intended to enhance deep learning applications on Earth monitoring tasks. The dataset spans 15 tera pixels of global remote-sensing data, combining imagery from a diverse range of sources, including NEON, Sentinel, and a novel release of 1m spatial resolution data from Satellogic. Our dataset provides a wide spectrum of image data with varying resolutions, harnessed from different sensors and organized coherently into an accessible HuggingFace dataset[1] in parquet format. This data spans five years, from 2017 to 2022. Accompanying the dataset, we introduce EarthMAE, a tailored Masked Autoencoder, developed to tackle the distinct challenges of remote sensing data. Trained in a self-supervised fashion, EarthMAE effectively processes different data modalities such as hyperspectral, multispectral, topographical data, segmentation maps, and temporal structure. This model helps us show that pre-training on Satellogic data improves performance on downstream tasks. While there is still a gap to fill in MAE for heterogeneous data, we regard this innovative combination of an expansive, diverse dataset and a versatile model adapted for self-supervised learning as a stride forward in deep learning for Earth monitoring.

1. Introduction

The instrumental role of Earth monitoring in navigating and confronting the escalating challenges of climate change, natural disasters, and environmental issues cannot be overstated [31]. It is also increasingly playing a central role in agriculture [11] and city planning [22]. Traditional vision models have already provided significant value across myriad applications [16,27,30], and the emergence of foundation models [13, 24, 25] promises to bring Earth monitoring to new horizons.

Today's large vision models, while proficient in detection [29] and image segmentation [8], are largely designed for RGB images from a first-person perspective. In contrast, remote sensing data offers unique properties: a **sky view** with a wide range of spatial resolutions, **multi-modality** including multispectral, radar, hyperspectral, point clouds, and elevation maps, and **temporality** from revisiting the same locations multiple times. Essentially, data with geographic coordinates can be matched with other data from the same location, providing an ever-expanding source of structured data suitable for large-scale self-supervised learning algorithms.

However, despite the surfeit of data that exceeds our current algorithmic processing capabilities, a significant portion of it remains out of reach, locked behind expensive paywalls. Free data sources such as Sentinel-1 and Sentinel-2, while useful, come with limitations: i) **low spatial resolution** of 10m of ground sample distance (GSD) and ii) **download difficulties** due to bandwidth throttling on Google Earth Engine and the cost associated with AWS for large-scale downloads.

To bridge this gap, we have teamed up with Satellogic and NEON to release a 15 tera pixels comprehensive, large-scale dataset designed specifically for self-supervised learning of extensive Earth monitoring algorithms. The dataset is available at Hugging Face and is conveniently partitioned to allow working on subsets of the data. This robust dataset comprises structured data derived from three distinct sources:

- Satellogic: Provides RGB and near-infrared data at 1m GSD, with temporal revisits and planet coverage.
- NEON: Provides 369 bands of hyperspectral data at 1m GSD, complemented by RGB data at 0.1m GSD and elevation data at 1m GSD across various US forests.
- Sentinel: We gathered a large structured subset of Sentinel-1 and 2, combining multispectral, synthetic aperture radar (SAR), and temporality.

The key contributions of this work are twofold:

- The introduction of a large-scale, multi-modal dataset tailored specifically for self-supervised learning in Earth monitoring.
- The development of a large masked auto-encoder trained with various self-supervision schema, demonstrating high performance across various Earth monitoring tasks.

2. Related Work

2.1. Datasets for training

The success of large-scale deep learning models has triggered research on larger datasets that can fit the capacity of current systems. [34] introduced BigEarthNet, a large-scale benchmark archive for remote sensing image understanding.

[1]Available at https://huggingface.co/datasets/satellogic/EarthView

This dataset consists of 590,326 Sentinel-2 image patches, annotated with multiple land-cover classes. The annotations were provided by the CORINE Land Cover database, and the dataset was significantly larger than existing archives in remote sensing. The authors demonstrated that training models on BigEarthNet improved accuracy compared to pre-training on ImageNet, indicating its potential for advancing operational remote sensing applications. [26] addressed the need for multi-label annotated datasets in remote sensing for semantic scene understanding. They developed MLRSNet, a multi-label high spatial resolution remote sensing dataset containing 109,161 samples within 46 scene categories. Each image in MLRSNet has at least one of 60 predefined labels, enabling training deep learning models for multi-label tasks such as scene classification and image retrieval. The authors highlighted the importance of MLRSNet as a benchmark dataset and its complementary nature to existing datasets like ImageNet. [37] presented the Five-Billion-Pixels dataset, aiming to enable country-scale land cover mapping with meter-resolution satellite imagery. The dataset comprises more than 5 billion labeled pixels from 150 high-resolution Gaofen-2 satellite images. They proposed a deep-learning-based unsupervised domain adaptation approach to transfer classification models trained on labeled data to unlabeled data for large-scale land cover mapping. The experiments demonstrated promising results across different sensors and geographical regions, showcasing the potential of the dataset and proposed approach. In a concurrent work, [2] introduced Satlas, a large-scale dataset for remote sensing image understanding. Satlas is comprehensive in terms of both breadth and scale, containing 302 million labels across 137 categories over a cumulative of 17 trillion pixels. The authors evaluated multiple baselines and a proposed method on Satlas. Pre-training on Satlas significantly improved performance on downstream tasks compared to ImageNet and other baselines. Lastly, the Umbra Open Data Program [1] features over twenty diverse time-series locations that are updated frequently, allowing users to experiment with SAR's capabilities.

While the previous benchmarks constitute a significant step in data availability for remote sensing, they are typically limited by the cost of obtaining labels. This has motivated the construction of unlabeled datasets that can leverage uncurated data from many different sources. For example, [18] proposed to leverage unlabeled data with Seasonal Contrast (SeCo). They collected a dataset of Sentinel-2 patches without human supervision, consisting of 1 million multispectral image patches from approximately 200,000 locations worldwide. By capturing seasonal changes with images from different dates, they aimed to enhance the training of models for remote sensing tasks. In a similar fashion, [33] proposed SSL4EO-L, consisting of 5 million unlabeled image patches from Landsat across 250,000 locations and multiple seasons. While SeCo and SSL4EO-L focused on uniformly covering most of the inhabited regions of Earth from a single data source, [4] focused on densely covering Europe with multiple data sources (Copernicus, Sentinel-2, and Planet) over space and time (500,000 locations in Europe with daily readings for a year) while [32] used EnMAP as their single source. In this work, we combine multiple data sources at different points in time while considering most of the inhabited Earth, resulting in a dataset that we named EarthView. Concretely, EarthView offers a larger and more diverse collection of unlabeled data by combining a high-quality curated selection from multiple data sources (Sentinel, NEON, and Satellogic), achieving a larger scale and variety than previous works (over 15 trillion pixels, with temporal revisits, and from 60 to 0.1m resolution). We share EarthView in a highly accessible format and available through Hugging Face, which enables easy integration into research projects. These qualities make our dataset a valuable resource for exploring uncharted patterns and structures in an unsupervised learning setting.

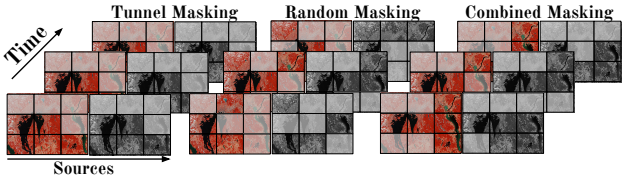

Figure 1. Different masking schemes explored in our work. Random masking, masks random patches across sources/time while tube masking masks the same patches. Combined masking combines both by first masking some patches consistently across sources/time and then randomly masking a subset of the remaining ones.

2.2. Learning from unlabelled data

Multi-view self-supervised learning methods have played a crucial role in building large models with remote sensing data [3, 18, 41]. In addition to multi-view SSL, reconstruction-based SSL with MAEs [9] has also been explored in the context of remote sensing. [40] introduced MIM, using masked image modeling for remote sensing scene classification. SatMAE [6] introduced a pre-training framework leveraging temporal and multispectral satellite imagery, encoding groups of bands independently with a spectral positional encoding. Scale-MAE and SatMAE++ additionally leverage information from multiple scales [23, 28]. SpectralMAE [44] and DOFA [42], focused on the reconstruction of arbitrary combinations of bands and data sources. Given the simplicity and versatility of MAE-based approaches to handling multiple data sources, we choose the SpectralMAE model class to experiment with the EarthView dataset introduced in this work. Concretely, we generalize SatMAE and SpectralMAE by combining multiple masking strategies, i.e. we combine masking all bands given a random position in an image with randomly masking individual bands in random positions (see Fig. 1). In experiments, we find that this strategy is effective for learning from heterogeneous data sources like the proposed EarthView data.

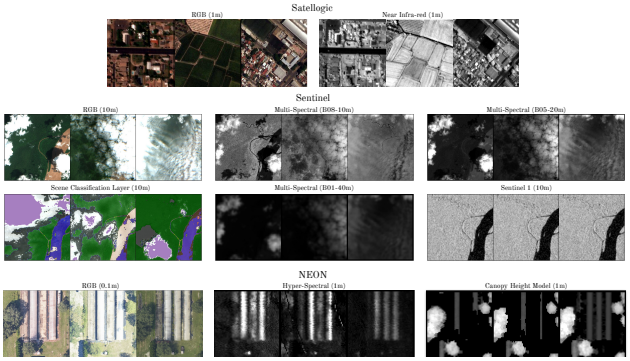

Figure 2. Samples from the dataset

3. Dataset

This work introduces an extensive dataset tailored for self-supervision on Earth monitoring data. It is based on the assumption that structure in the data brings an essential signal to self-supervised algorithms for finding high-level semantic representation that can make sense of the data. To this end, we combine spectral, synthetic-aperture radar (SAR), temporal, and spatial structures in a large-scale dataset composed of different sources and multiple spatial resolutions. The data gathered for this project is drawn from a triad of distinct sources, namely Satellogic, Sentinel and NEON (Fig. 2). Each of these contributes unique facets and dimensionalities to the integrated dataset.

3.1. Satellogic

Vision Satellogic is a provider of high-resolution remote sensing imagery, whose main mission consists of democratizing access to Earth Observation Data. To date, Satellogic's fleet consists of 24 active satellites, having successfully put over 50 satellites in orbit in the past years. These satellites provide high-resolution imagery that ranges from 50cm to 1m.

Building multimodal foundation models is paramount to further advance Earth Observation applications. Thanks to these models, the EO community can easily construct on top of the new solutions without the need to collect large amounts of data. Multimodal foundation models, such as vision-language large models, to be effective require to be trained on vast amounts of data. More importantly, in this specific case, data should cover a large variety of regions over the planet to have the most accurate representation of Earth. To this end, Satellogic offers to release a portion of its second half of 2022, ranging from the 1st of July to the 30th of December.

Accessibility High-resolution data covering different regions of the world is not freely accessible. To truly push the community towards building open foundation models, Satellogic releases its dataset under the license CC BY 4.0. To our knowledge, this release represents one of the largest contribution to the EO community ever made by a private company. Moreover, this is the first one that includes the visual and top-of-atmosphere products consisting both of four bands (red, green, blue and near-infrared) at a 1-meter resolution. Satellogic also offers metadata such as the off-nadir angle, sun elevation, azimuth angle, or timestamp. The Satellogic dataset stands out as unique compared to other publicly available datasets for several reasons:

- It covers diverse regions of the Earth, unlike **NAIP** [39], **LINZ** [15], or **NEON**
- It offers high-resolution imagery, surpassing traditional open data sources like **Sentinel-2** and **Landsat 8 and 9**
- It allows for commercial use, as opposed to datasets such as **DOTA** [7], **UC Merced Land-Use** [43] or **xView** [14]
- It includes rich metadata and geolocation
- Its size is multiple orders of magnitude larger than most of the publicly available datasets, *e.g.* **FAIR1M** [36], **DIOR** [17], **NWPU-RESISC-45** [5] and **Floodnet** [35]

Sensor Imagery is acquired at 1m GSD from space, at different off-nadir angles, over four bands (blue, green, red and near-infrared).

Spatial Distribution The acquisition of imagery during the 2nd half of 2022 was performed on demand by multiple customers. As can be seen in Fig. 3, the Satellogic dataset covers different regions over all the continents. To build these regions, we make use of 3,758 captures, where all these captures have a percentage of clouds below 30%. Out of these captures, we create unique patches. In particular, for each capture, we start at the top-left position and crop non-overlapping windows of 384×384 pixels. The dataset comprises a total of 2,967,663 patches, summing up 437,682 km^2. If we discard overlapping regions[2], the dataset covers a 10% smaller area (396,280 km^2). Most of these overlapping regions have less than 50% redundancy and are comparable to data augmentation.

Temporality The resulting set of patches contains a varying number of revisits, depending on how many captures were performed over the same area. These revisits range from 1 to 68, where 986,521 regions have at least two revisits. The almost 3M patches translate to 6,165,992 images including revisits.

3.2. Sentinel

Accessibility While Sentinel data is distributed under a Creative Commons license, very large datasets are less accessible. Since the Google Earth Engine (GEE) throttles the download speed, it is prohibitively long to download terabytes of data. We thus had to resort to AWS, but since *requester pays* the bandwidth, it still required a large budget just to collect this data.[3]

Sensors Sentinel's satellites offer a wide range of sensors. For this project, we focus on SAR from **Sentinel-1**, and

[2] An artifact of our patching algorithm

[3] AWS stores data in large tiles, requiring many downloads for broad coverage despite needing only fractions.

multispectral from **Sentinel-2**. The other sensors offer a spatial resolution that is too low for our purpose. For SAR, we use the level-1 Ground Range Detected (GRD) product available in AWS. We stack the different polarizations (VH, VV) resulting in two bands, and resample it to 10m resolution. Sentinel-1 images are then saved in uint16 format to reduce their size in bytes. Finally, Sentinel-2 is composed of 13 spectral bands. The main bands, (blue, green, red, near-infrared), have 10m GSD, but due to atmospheric absorption of other wavelengths other bands have 20m and 60m resolution.

Spatial Distribution Our aim is to gather a wide range of regions covering the planet, however, we also want to avoid highly redundant patterns such as ocean, desert, and forests. To this end, we gather inspiration from [18] and collect Sentinel-2 tiles that overlap regions within a 50km radius around the top largest cities in the world. Each footprint area (100km x 100km) is large enough to cover coastal, agricultural and rural regions. We also sample Sentinel-2 tile regions that cover Satellogic data. Since Sentinel-1 does not follow the same grid system as Sentinel-2, we use the collected Sentinel-2 tile footprints to query Sentinel-1 captures, and crop them accordingly.

For each Sentinel-2 tile footprint, we extract 500 non-overlapping regions of $3{,}840$ m x $3{,}840$ m. Out of all possible candidates (over 670) per tile, we select the best ones based on the cloud coverage and the entropy of Sentinel-2's scene classification layer (to increase diversity within patches). We end up collecting over 2,000 Sentinel-2 tiles, resulting in over 1M unique regions. These same regions are used to build the Sentinel-1 collection. Similar to Satellogic, some of the 1M regions do overlap. The 1M regions sum up a total of $15{,}474{,}873$ km^2, if we take into account the overlapping areas, the coverage area reduces to $15{,}074{,}640$ km^2. Out of the 1M regions, there are $65{,}251$ that overlap, covering $388{,}840$ km^2. A large part of these overlapping regions have less than 50% redundancy.

Temporal distribution Temporality also offers an important signal for a model to learn how scenes evolve over time. However, a long sequence could significantly increase the redundancy and size of the dataset. Hence we limit to ten revisits per location, where six are densely sampled over time and the other four are sampled with three-month intervals to ensure coverage of the seasons. At the same time, during the selection process, we check multiple time sequences under these same constraints. For each sequence, we extract the percentage of clouds for all the dates and keep it as an indicator of how good the sequence is. Among all the sequences, we select the best candidate. This results in sequences ranging from 2017 to 2022. For Sentinel-1, we did not have access to the same temporality. During the collection process, we choose the closest dates to the ones we have for Sentinel-2. When feasible, we also select other dates available within Sentinel-2 range to increase the number of Sentinel-1 revisits.

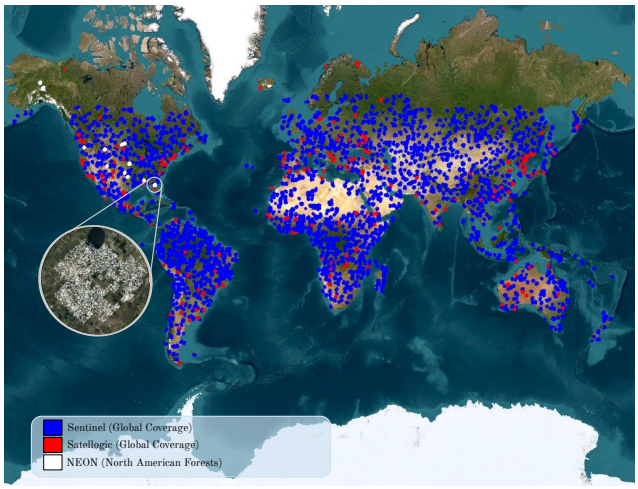

Figure 3. Spatial coverage for each source. Note that a colored area may contain multiple patches.

3.3. NEON

NEON data combines high-resolution RGB, hyperspectral, and lidar data for the study of ecological sites in the United States.

Accessibility NEON data is redistributed under CC0 1.0 and accessible on the NEON data portal.

Sensors NEON offers high-resolution RGB at 0.1m [20] GSD and hyperspectral data [21] comprised of 426 spectral bands at 1m GSD. It is also accompanied by lidar, which is post-processed to estimate the tree canopy height at 1m GSD [19].

Spatial Distribution This incredibly high-resolution data comes with very limited spatial distribution. We have collected data from 12 of the available sites with multiple sub-locations on each of these sites (See Fig. 3). Each location covers 64m $\times$ 64m, where depending on the sensor, we have 640 pixels $\times$ 640 pixels or 64 pixels $\times$ 64 pixels. All locations sum up 148.76 km^2. Similar to Satellogic and Sentinel, Neon has some redundancy. The overlapping regions cover less than 5%, where most of these overlapping regions have less than 50% redundancy.

Temporality This data also offers yearly revisits with some limitations. Most sites contain a maximum of 3 revisits and the exact date was not collected. Nevertheless, we matched all available revisits for each location that we collected.

3.4. Hosting and Storage

Hosting Through a partnership with HuggingFace, data is made available at their servers.This offers a free download and broad access to the whole community. For other commercial cloud providers, even if hosting were free,

Table 1. Dataset overview

	Sensor	# bands	GSD (m)	Pixel per patch	area (m)	# revisits	# patches	# Giga gray pixels
Satellogic	RGBN	4	1	384 x 384	384 x 384	1 - 5	2967663	3636.85
Sentinel-1	SAR	2	10	384 x 384	3840 x 3840	3 - 9	1049466	1743.61
Sentinel-2	MS	13	10, 20, 60	384 x 384	3840 x 3840	10	1049466	9086.36
NEON-RGB	RGB	3	0.1	640 x 640	64 x 64	3	35501	130.87
NEON-Hyper	HS	369	1	64 x 64	64 x 64	3	35501	160.97
NEON-Elev	Lidar	1	1	64 x 64	64 x 64	3	35501	0.44

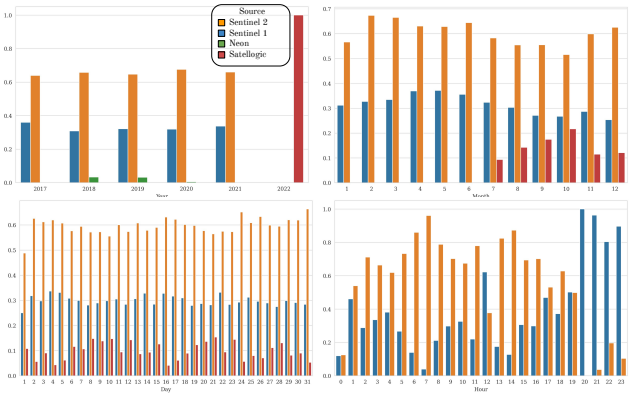

Figure 4. Temporal distribution of the dataset. NEON data only provides the year, and Satellogic data does not contain the time of the day.

downloading 15 TB of data as *requester pays* would cost on the scale of 1000 USD per download.

Storage For accessibility and to minimize bandwidth, we store the dataset in functional subsets. That is, each of the different sensors can be downloaded separately. Moreover, each subset is split into shards and these can be downloaded independently, this allows the user to download only a set of the data if needed. Each shard contains a series of regions.

Format Each region is represented as a dictionary containing the sensor data with the RGB data along with the other bands grouped by resolution (e.g., RGB, 10m, 20m, etc) arrayed in a four-dimensional matrix structure (time, spectral bands, height, width). The dictionary also contains the metadata available for the region (e.g., bounds, timestamps, etc). As metadata, we provide the geo-referenced bounding box and some form of timestamp see (Fig. 4) for all sensors. However other metadata, such as solar angles and incidence angles is only available for Sentinel-2.

4. Model

In this work, we leverage a Masked Autoencoder (MAE) [10], distinguished by its asymmetrical encoder-decoder architecture. It incorporates an encoder that functions exclusively on a visible subset of patches, and a streamlined decoder, that rebuilds the original image from the latent representation and mask tokens. This model, recognized for its proficiency in self-supervised learning tasks, has been appropriately restructured to manage remote sensing data as described next.

4.1. EarthMAE

Our EarthMAE model (Fig. 5) remains faithful to the original architecture, albeit we adjusted the tokenizers and positional encodings to leverage time and different modalities.

Tokenizers We incorporated a distinct tokenizer (linear transform) for each source, owing to the fact that different sources contain a disparate number of channels. This method offers a more nuanced comprehension of the data, accounting for the varied characteristics associated with different bands and sources. To ensure that all bands and sources produce the same amount of patches we resize all images to 224×224.

Encoding Analogous to positional encodings, we introduce source and temporal encodings to capture the unique characteristics of the data. The source encoding is generated by embedding the source labels into a vector space, allowing the model to differentiate between various data sources such as multispectral, RGB, and hyperspectral images.

For the temporal encoding, we leverage the timestamp metadata provided in our dataset. We decompose each timestamp into discrete components: *year*, *month*, *day*, and *hour*. Each component is mapped to a 16-dimensional embedding vector using separate embedding layers, each initialized randomly. Specifically, we use embedding layers with sizes: 7 for year (representing six possible years plus an index for unknown), 13 for month, 32 for day, and 25 for hour, to accommodate all possible values and account for unknown timestamps. The embeddings for the time components are concatenated to form a 64-dimensional time embedding for each timestep.

Formally, given a batch size B, number of timesteps t, number of sources s, and number of patches p, we process the timestep tensor of shape $(B,\ t,\ 4)$, where the last dimension corresponds to [year, month, day, hour]. We apply the respective embedding layers to each time component, obtaining embeddings of shape $(B,\ t,\ 16)$ for each. These are concatenated along the last dimension to form the time embedding tensor of shape $(B,\ t,\ 64)$. The time embeddings are then expanded and combined with the positional encodings and source embeddings.

The source embeddings are generated by applying an

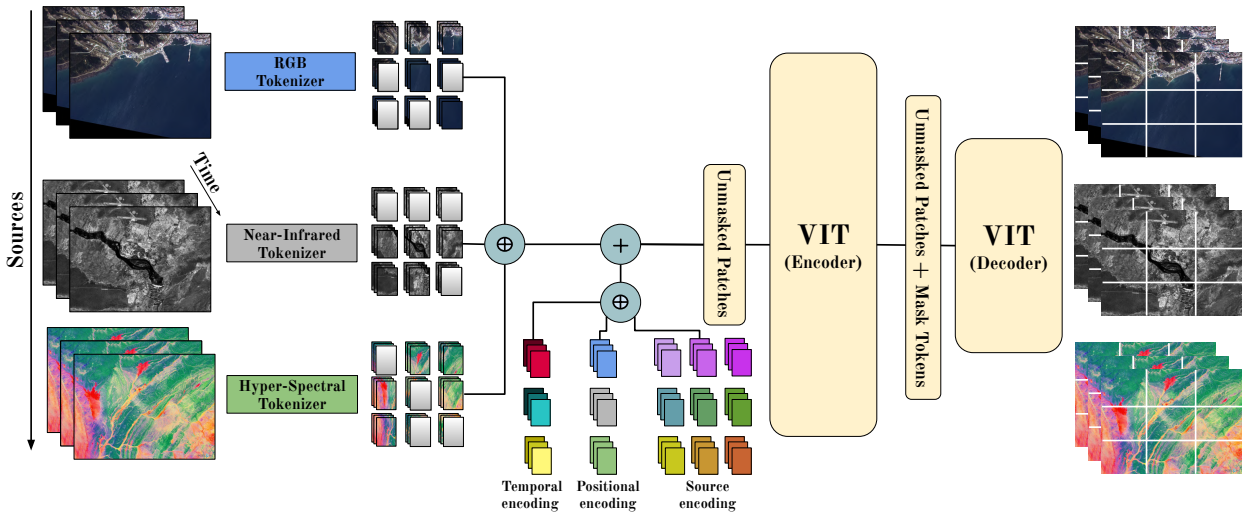

Figure 5. EarthMAE: The model leverages time information and can digest data from an arbitrary number of sources. Each input source is tokenized into a fixed number of patches and then all patches are concatenated. The time, source, and positional encodings are concatenated and added to the patches.

embedding layer to the source labels, resulting in a tensor of shape (B, s, d) with $d = 64$. After expanding and aligning dimensions, the time embeddings, positional encodings $(B,\ p,\ D)$, and source embeddings are concatenated along the feature dimension to form the final encoding tensor of shape (B, t, s, p, D'), where $D' = D + d + 64$.

This comprehensive encoding allows the model to incorporate temporal, positional, and source information, enhancing its ability to process data with multiple sources and variable timesteps, which is essential for practical remote sensing tasks. Note that since downstream tasks may not always have time information, we simulate missing timestamps in the data by occasionally setting the timesteps to zero vectors, representing unknown time components. Specifically, with a probability of 10%, we replace the actual timesteps with zeros during training:

4.2. Training Paradigm

Our training approach uses a task-based system, and diverse masking strategies, and includes temporal information. These components work together to improve the flexibility and effectiveness of our EarthMAE model. This comprehensive approach matches well with the challenges presented by the various sensor data, timesteps, and masking schemes that are typical in the self-supervised learning of remote sensing data.

The training approach makes the most of the unique mix of data in our dataset. This includes varied sensors and data types, such as multispectral data from Sentinel, hyperspectral data from NEON, RGB data from Satellogic, and specific bands like Sentinel-2 RGB and SCL used for segmentation tasks. The training paradigm can be divided into tasks where each task is essentially a subset of the sources available in the data (e.g., RGG, CHM, etc), this simplifies the handling and combination of data sources of different channels, bands, sizes, and resolutions. To manage this broad spectrum of data, we have set up a task-based training system. Here, we distribute tasks across multiple GPUs, with each GPU handling a specific task. This way, we can process different types of data and sensors in parallel, making the process more efficient.

Loss The training objective is the mean squared loss (MSE) between the model's reconstruction and the normalized pixel values on masked patches same as in the original MAE [9] paper.

Masking Simply performing random masking such as in the original MAE can lead to very easy reconstruction tasks where the model can learn to translate from one source to another or copy most of a patch from another time step. To ensure challenging reconstruction tasks with video sequences, [38] proposes tube masking where a given patch is masked across all time steps to prevent *information leak* through time. We incorporated tube masking in our implementation.

Our experiments aimed to understand the impacts of different data sources (NEON, Sentinel, Satellogic), the inclusion of temporality, and various masking strategies on the performance of our Masked Autoencoder (MAE) model.

4.3. Experimental Setup

To evaluate each pre-trained model, we leverage the classification benchmark of GeoBench [12]. This benchmark is specifically designed to evaluate pre-trained models on remote sensing data. They curated 6 classification datasets: m-BigEarthNet, m-Brick-Kiln, m-EuroSAT, m-ForestNet, m-PV4GER, and m-SO2SAT, to cover a range of downstream tasks. On each downstream task, the pre-trained model is fine-tuned over parameters obtained through random search,

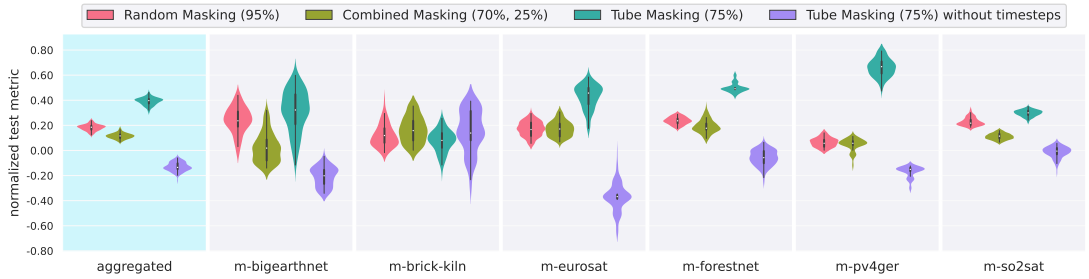

Figure 6. Performance of different masking schemes. Tube masking with a 75% ratio consistently outperforms the rest. Combined masking refers to tube masking 75% of the patches and randomly 25% of the remaining ones. Pre-training without time information (purple) hurts performance. Results are reported across 5 different seeds.

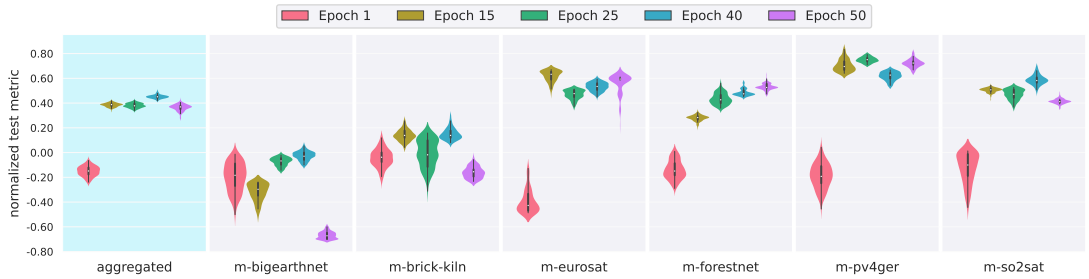

Figure 7. Performance on downstream tasks. Performance metrics are normalized and plotted for different epochs of model pre-training: Epochs 1, 15, 25, 40, and 50. The model shows varied performance across tasks, with some showing rapid improvement as early as Epoch 15. For example, m-forestnet and m-pv4ger exhibit a noticeable upward trend, suggesting that these tasks benefit early from the features learned during pre-training. Results are reported across 5 different seeds.

and the best hyperparameter is selected on the validation set and re-trained with 5 different seeds to be evaluated on the test set. See supplementary material for training and fine-tuning details. A bootstrap procedure is used to report the uncertainty of the interquartile mean[4]. A single aggregated result is obtained by averaging the normalized score[5].

Following the standard MAE training process, all models were pre-trained for 100 epochs using a 90% masking ratio with tube masking [38], where all timesteps and sources are masked in the same way (tube masking). We also experimented with different masking strategies by introducing random masking (where timesteps and sources are masked randomly with a 95% ratio) and combined masking, which mixes tube and random masking strategies.

4.4. Results and Discussion

Our results showed significant variations in model performance, depending on the data sources, timesteps inclusion, and the applied masking schemes.

Fig. 6 shows the performance for models trained on Sentinel data. Tube masking outperforms the rest in most downstream tasks while the removal of time information from the pre-training hurts performance in all datasets. For the results in this work, we use tube masking schema with a 75% masking ratio, unless specified otherwise.

In Fig. 7, we verify the convergence behavior of our training process, when training for 50 epochs with cosine annealing of the learning rate. The figure depicts the downstream performance for 5 checkpoints along the 50 epochs. The aggregated results show a mostly stable performance after 15 checkpoints, where the variability is likely due to the cross-checkpoint variance. Surprisingly, on certain tasks like m-bigearthnet, we observed an impressive drop in performance, that we were not able to explain. On the other hand, some tasks, like m-forestnet, expose a good progression of performance.

As seen in Fig. 8, pre-training on Satellogic data offers consistent performance improvement over models pre-trained on Sentinel data alone, with the model that combines both data sources outperforming the rest. This suggests that the high resolution of Satellogic offers a useful signal during pre-training.

5. Experiments

Aggregated results in Fig. 9 show that the performances of this model are significantly lower than other pre-trained models evaluated in [12]. Hypothesis for such discrepancies could be that MAE is not the right training loss or that remote sensing data alone is insufficient or too redundant.

Fig. 10 shows how performance on downstream tasks varies for different dataset sizes. For harder tasks, the

[4]Average discarding top and bottom 25% to reduce outliers.

[5]Normalization constants set weak baselines (e.g., ResNet18) to 0 and strong baselines (e.g., SwinV2) to 1.

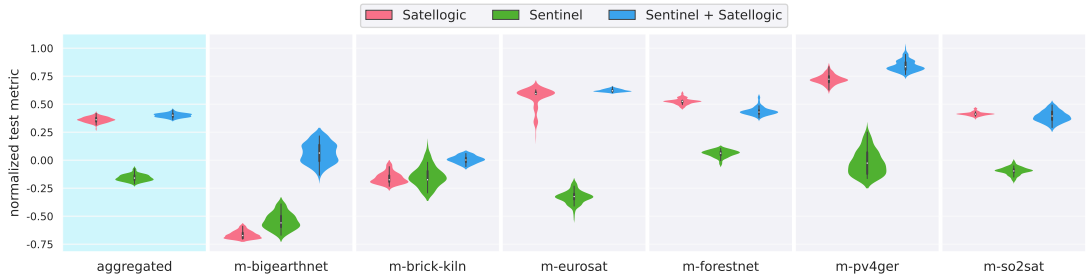

Figure 8. Downstream task performance for models pre-trained on different data. Note that including Satellogic data, whether alone or combined with Sentinel data, consistently enhances the model's performance across all tasks compared to using only Sentinel data. The combination of Sentinel and Satellogic data achieves the highest performance improvements. Results are reported across 5 different seeds.

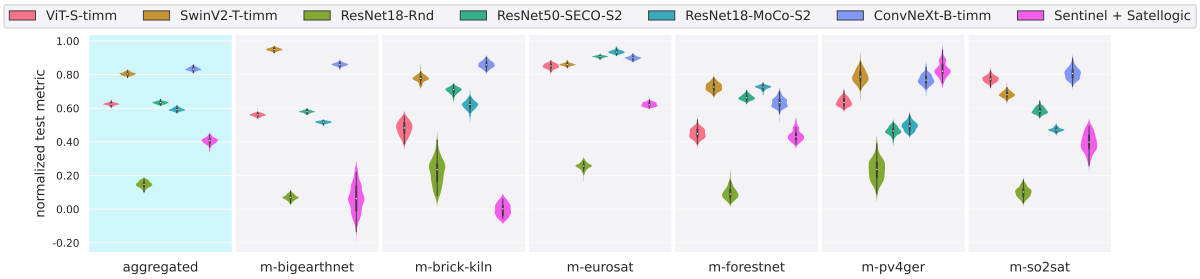

Figure 9. Downstream task performance of an MAE model pre-trained on Sentinel and Satellogic data and models featured on GeoBench [12]. Results are reported across 5 different seeds.

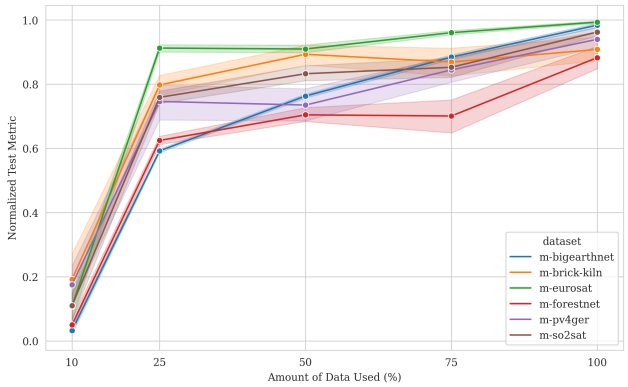

Figure 10. Performance on downstream tasks for different dataset sizes shows that for some tasks, performance increases almost linearly, indicating considerable room for improvement with larger datasets.

performance increases almost linearly indicating that despite the potential redundancy of satellite imagery, our dataset is diverse enough so that using it at full scale offers improved results. It seems that for tasks like m-bigearhtnet and m-forestnet would benefit from an even larger scale.

6. Conclusion

This paper primarily focuses on the introduction and description of our unique and expansive remote sensing dataset. With multiple sensors and data types, it provides researchers and practitioners with a wide array of information, which is anticipated to significantly advance the field of remote sensing and related applications.

While we have also presented EarthMAE, a tailored model designed to handle the nuances of our dataset, the overarching theme of our work is the sheer potential held within the dataset itself. The dataset's size and diversity enable an exhaustive examination of different sensor types and data structures. Furthermore, the distribution of tasks across multiple GPUs during training fosters an efficient environment for exploring various self-supervised learning scenarios.

In our investigation, we explored different masking strategies each of which holds implications for the model's performance. We also incorporated temporal information from the dataset into the model using timestamp metadata, a strategy expected to increase accuracy across different remote sensing tasks.

Ultimately, the value of this dataset extends beyond the scope of our work. It provides an open playing field for future explorations in self-supervised learning and remote sensing applications. We hope that the efforts encapsulated in this paper serve as a springboard for future research.

Limitations While EarthView offers diverse sources, sensors, and scales, it lacks modalities like text. The EarthMAE model does not fully realize EarthView's potential. Researchers are encouraged to explore larger models trained on this dataset with others like [2].

References

[1] Umbra synthetic aperture radar (sar) open data. https://registry.opendata.aws/umbra-open-data. Accessed on January 11, 2025. 2

[2] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlas: A large-scale, multi-task dataset for remote sensing image understanding. *arXiv preprint arXiv:2211.15660*, 2022. 2, 8

[3] Jasmine Bayrooti, Noah Goodman, and Alex Tamkin. Multispectral self-supervised learning with viewmaker networks. *arXiv preprint arXiv:2302.05757*, 2023. 2

[4] Priyash Bhugra, Benjamin Bischke, Christoph Werner, Robert Syrnicki, Carolin Packbier, Patrick Helber, Caglar Senaras, Akhil Singh Rana, Tim Davis, Wanda De Keersmaecker, et al. Rapidai4eo: Mono-and multi-temporal deep learning models for updating the corine land cover product. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 2247–2250. IEEE, 2022. 2

[5] Guang Cheng, Jun Han, Xin Li, Liangpei Zhang, Jun Liu, and Lihong Zhao. Nwpu-resisc45: A real-world remote sensing image scene classification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 150–158. IEEE, 2017. 3

[6] 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. 2

[7] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 1

[9] 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. 2, 6

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B Girshick. Masked autoencoders are scalable vision learners. corr abs/2111.06377 (2021). *arXiv preprint arXiv:2111.06377*, 2021. 5

[11] Sami Khanal, Kushal Kc, John P Fulton, Scott Shearer, and Erdal Ozkan. Remote sensing in agriculture—accomplishments, limitations, and opportunities. *Remote Sensing*, 12(22):3783, 2020. 1

[12] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: Toward foundation models for earth monitoring, 2023. 6, 7, 8

[13] Alexandre Lacoste, Evan David Sherwin, Hannah Kerner, Hamed Alemohammad, Björn Lütjens, Jeremy Irvin, David Dao, Alex Chang, Mehmet Gunturkun, Alexandre Drouin, Pau Rodriguez, and David Vazquez. Toward foundation models for earth monitoring: Proposal for a climate change benchmark, 2021. 1

[14] Dominic Lam, Kathleen Lejeune, Joel Farrell, John Tighe, Travis Chapman, Paul Corcoran, Daniel Tingdahl, Nikhil Manohar, Vu Nguyen, Yangxiaokang Hwang, et al. xview: Objects in context in overhead imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2018. xView Dataset. 3

[15] Land Information New Zealand. LINZ Data Service, 2024. Accessed: 2024-08-29. 3

[16] Issam Laradji, Pau Rodriguez, Freddie Kalaitzis, David Vazquez, Ross Young, Ed Davey, and Alexandre Lacoste. Counting cows: Tracking illegal cattle ranching from high-resolution satellite imagery. *arXiv preprint arXiv:2011.07369*, 2020. 1

[17] Zhangjie Li, Hengrong Li, Xue Huang, Wei Hu, Shuguang Wang, and Sheng Li. Dior: A large-scale dataset for object detection in remote sensing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4118–4127. IEEE, 2018. 3

[18] 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, 4

[19] NEON (National Ecological Observatory Network). Ecosystem structure (dp3.30015.001), provisional data, 2024. Dataset accessed from https://data.neonscience.org/data-products/DP3.30015.001 on October 29, 2024. Data archived at [your DOI]. 4

[20] NEON (National Ecological Observatory Network). High-resolution orthorectified camera imagery mosaic (dp3.30010.001), provisional data, 2024. Dataset accessed from https://data.neonscience.org/data-products/DP3.30010.001 on October 29, 2024. Data archived at [your DOI]. 4

[21] NEON (National Ecological Observatory Network). Spectrometer orthorectified surface directional reflectance - mosaic (dp3.30006.001), provisional data, 2024. Dataset accessed from https://data.neonscience.org/data-products/DP3.30006.001 on October 29, 2024. Data archived at [your DOI]. 4

[22] Maik Netzband, William L Stefanov, and Charles Redman. *Applied remote sensing for urban planning, governance and sustainability*. Springer Science & Business Media, 2007. 1

[23] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. *arXiv preprint arXiv:2403.05419*, 2024. 2

[24] OpenAI. Gpt-4 technical report, 2023. 1

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien

Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[26] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 2

[27] 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. 1

[28] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2022. 2

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[30] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M Gonfaus, Kamal Nasrollahi, Thomas B Moeslund, and F Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, 52(5):3314–3324, 2017. 1

[31] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022. 1

[32] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2175, 2023. 2

[33] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[34] 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

[35] Ananya Swaminathan, Abhinav Gupta, Prem Natarajan, Nikhil Sinha, Pallabi Das, et al. Floodnet: A multimodal dataset for flood event detection and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243. IEEE, 2021. 3

[36] Zhi Tang, Wen Li, Jian Liu, Jiahui Yang, Shiyu Zhao, Shizhe Yang, Di Wu, Xiang Zheng, and Fuchao Wu. Fair1m: A benchmark dataset for fine-grained aerial image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12287–12296. IEEE, 2020. 3

[37] Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:178–196, 2023. 2

[38] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*. 6, 7

[39] U.S. Department of Agriculture, Farm Service Agency. NAIP Imagery, 2021. Accessed: 2024-08-29. 3

[40] Liya Wang and Alex Tien. Remote sensing scene classification with masked image modeling (mim). *arXiv preprint arXiv:2302.14256*, 2023. 2

[41] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. *arXiv preprint arXiv:2303.06670*, 2023. 2

[42] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024. 2

[43] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010. UC Merced Land Use Dataset. 3

[44] Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7):3728, 2023. 2