APPENDIX: MULTIMAE MEETS EARTH OBSERVATION: PRE-TRAINING MULTI-MODAL MULTI-TASK MASKED AUTOENCODERS FOR EARTH OBSERVATION TASKS

Author(s) Name(s)

Author Affiliation(s)

Contents

1. DATA DETAILS

1.1. Sentinel-2 data

Band	Description	Resolution	Wavelength (nm)
B1	Ultra blue (Aerosol)	60	443
B2	Blue	10	490
B3	Green	10	560
B4	Red	10	665
B5	Red edge 1 (near infrared)	20	705
B6	Red edge 2 (near infrared)	20	740
B7	Red edge 3 (near infrared)	20	783
B8	Near infrared	10	842
B8A	Red edge 4 (near infrared)	20	865
B9	Water vapor	60	940
B10	Cirrus	60	1375
B11	Shortwave infrared 1 (SWIR)	20	1610
B12	Shortwave infrared 2 (SWIR)	20	2190

Table 1. **Sentinel-2 bands details.** Details for each of the spectral bands composing sentinel-2 data [1, 2].

Sentinel-2 (S2) imagery comprises 13 spectral bands extending across the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions of the electromagnetic spectrum.

Name	Description	Data type	Bands	Used
Pixel-level modalities				
Sentinel-2	Optical	Continuous	13	✓
Sentinel-1	SAR	Continuous	8	×
Aster DEM	Elevation	Continuous	2	✓
ETH-GCHM	Vegetation height	Continuous	2	×
ESA World Cover	Landcover	Categorical	1	✓
Dynamic World	Landcover	Categorical	1	×
Image-level modalities				
Biome	Landcover	Categorical	1	×
Ecoregion	Landcover	Categorical	1	×
ERA5 temperature	Climate analysis	Continuous	9	×
ERA5 precipitation	Climate analysis	Continuous	3	×
Geolocation	Latitude, Longitude	Continuous	4	×
Date	Month of the year	Continuous	2	×

Table 2. **Details of modalities from MMEarth [3] dataset.** In this version of our approach, we strategically rely only on a subset of pixel-level (visual) modalities, as indicated by the last column of the table.

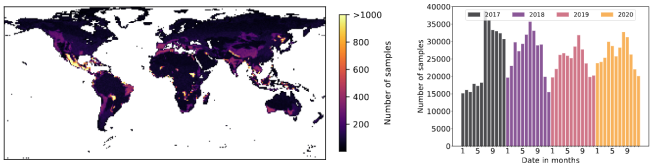
These bands are provided at three different spatial resolutions: four bands at 10 m, six bands at 20 m, and three bands at 60 m. The detailed characteristics of these bands are summarised in Table 1.

1.2. Pre-training data

For the pre-training stage, we rely on the MMEarth dataset [3]. It represents one of the most recent and complete multi-modal large-scale collections of EO data. MMEarth matches ImageNet-1k [4] size, containing 1.24 million samples. It comprises 12 aligned modalities distributed in two groups: pixel-level and image-level. The first group includes visual data, such as optical, SAR, landcover labels and elevation maps. The second group includes metadata, e.g., date, temperature information, and geolocation. Table 2 provides further details on the MMEarth dataset, while Figure 1 illustrates its spatial and temporal distribution.

1.3. Fine-tuning data

For fine-tuning, we utilise mostly data from GEO-Bench [6] datasets. This benchmark represents an effort to provide diverse data for fine-tuning pre-trained models on different



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

Fig. 1. Spatial and temporal distribution of MMEarth dataset. Data from MMEarth spans across 4 years from multiple world regions. Multi-modal data has been collected and properly aligned using Google Earth Engine Platform [5]. Figure taken from [3].

Name	Image Size	Classes	Train / Val / Test	Bands
		Classification tasks		
m-eurosat [6]	64×64	10	2k / 1k / 1k	13
m-brick-kiln [6]	64×64	2	15k / 1k / 1k	13
m-so2sat [6]	32×32	17	20k / 1k / 1k	18
m-bigearthnet [6]	120×120	43	20k / 1k / 1k	12
EuroSAT [7]	64×64	10	16.2k / 5.4k / 5.4k	13
fMoW (10%) [8]	64×64	62	71.3k / 85k / 85k	13
		Segmentation tasks		
m-SA-crop-type [6]	256×256	10	3k / 1k / 1k	13
m-cashew-plantation [6]	256×256	7	1.3k / 400 / 50	13

Table 3. EO datasets used for fine-tuning on downstream classification and segmentation tasks. Summary of datasets used for evaluating the transfer learning capabilities of our approach. Most datasets come from Geo-Bench [6] such as those indicated with the prefix *m-*. Other standard datasets like EuroSAT [7] and fMoW [8] are included for broader comparisons.

downstream EO tasks. GEO-Bench adheres to the following design principles that make it suitable for properly evaluating the transfer learning capabilities of EO models: ① Ease of use. ② Expert knowledge incorporation. ③ Diversity of tasks. ④ Original train, validation, and test splits. ⑤ Permissive license [6].

Overall, [6] comprises multiple modified versions of standard geospatial datasets for classification and segmentation tasks. We use a subset of those datasets as shown in Table 3. For fine-tuning on classification tasks, we add a couple of standard datasets used in previous related works: EuroSAT [8] and S2 version of fMoW [7] datasets, which allows for broader comparisons. According to [6], using small datasets aligns better with fine-tuning philosophy in the EO context. Thus, we reduce fMoW [8] and only utilise 10% of it. Apart from this exception, all the other data collections used for fine-tuning remain unmodified.

2. PRE-TRAINING MULTIMAE

2.1. Pre-training objective

We pre-train our approach (depicted in Figure 2) using six input modalities: RGB, IRED, SIRED, EB, DEPTH, and SEG. Four of them come from Sentinel-2 data. We use all available samples in the MMEarth dataset as indicated by subsection 1.2. We follow a self-supervised reconstruction pre-training objective similar to standard MAEs [9]. Following previous approaches [9, 10], we rely on a MSE (Mean Squared Error) loss on the reconstructed tokens. However, since our approach seeks to reconstruct various inputs via N separate decoders D_i, we average the individual reconstruction losses, as indicated by Equation 1,

$$\mathcal{L} = \sum_{i=1}^{N} MSE(D_i(x_m, x_a), \hat{x}_m) \tag{1}$$

where x_m and x_a correspond to the decoders inputs, i.e. modality-specific tokens and all modalities tokens, respectively, while $\hat{x}_m$ represents the ground truth tokens. In our case, N is set to 6 according to the number of input modalities.

2.2. Decoders design

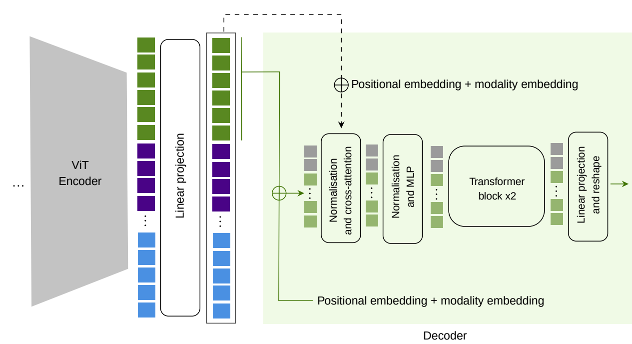


Fig. 3. Decoders design. The tokens from the encoder are firstly linearly projected to match the decoder dimension. Then, modality-specific and positional embeddings are added. A cross-attention layer incorporate information from tokens of the general representation of all the modalities, which is then processed by an MLP and a couple of transformer blocks. Finally, tokens are projected and reshaped to build an image.

Our decoders follow the design of those in previous works [10, 9]. Each decoder in our approach contains a linear projection layer that adapts the encoder's output to the decoder dimension. Then, after the linear projection, it adds to the decoder's inputs sine-cosine positional embeddings and the

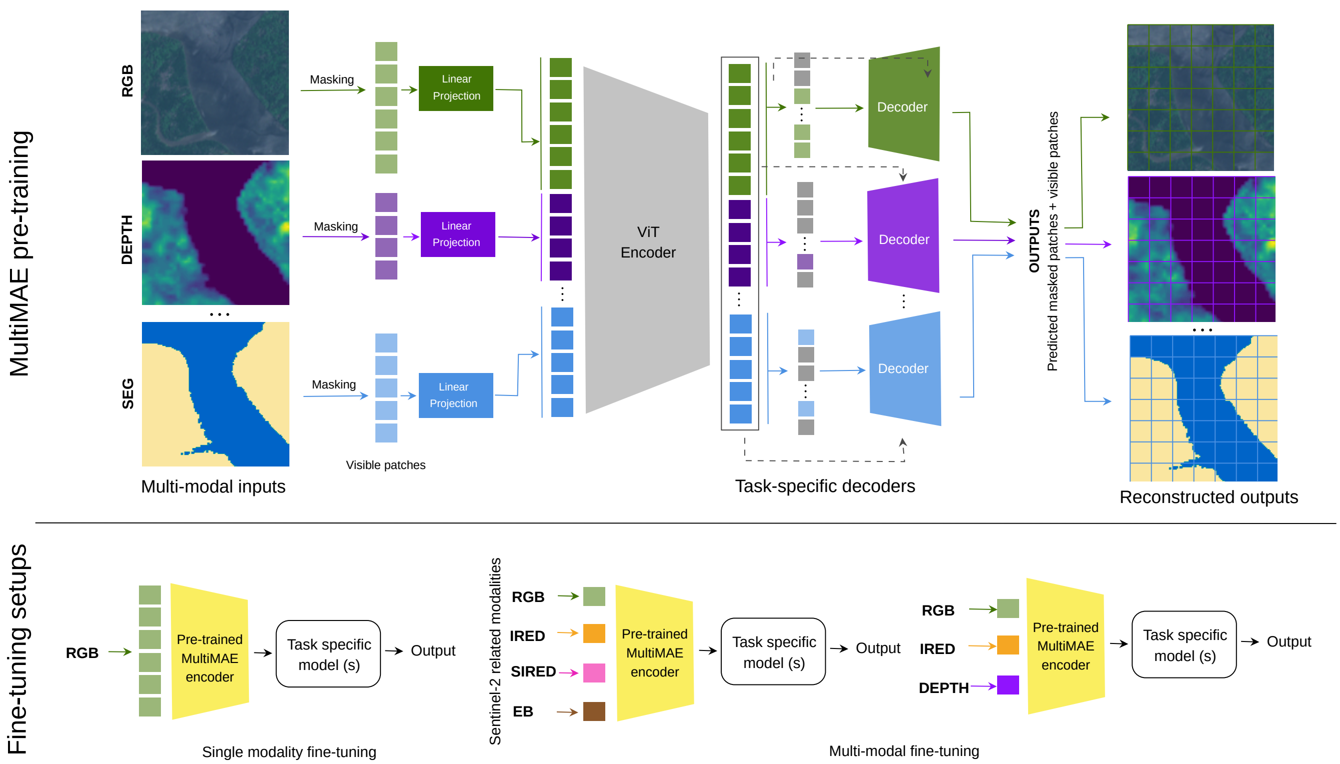


Fig. 2. **MultiMAE pre-training and fine-tuning with EO data.** The top part of the figure illustrates the pre-training stage with six input modalities from EO data: RGB, IRED, SIRED, EB, DEPTH, and SEG (for simplicity, only three are depicted in the figure). The bottom part depicts fine-tuning setups. When fine-tuning, task-specific models are coupled with a pre-trained MultiMAE encoder. Fine-tuning occurs under multiple scenarios, e.g. single-modality or multi-modality, by varying the number of input modalities.

learned modality embeddings. This is further processed by a cross-attention layer, an MLP, and two transformer blocks as illustrated by Figure 3. Using fewer transformer blocks in the decoders makes our approach computationally efficient.

3. FINE-TUNING SETUPS

For classification tasks, we couple the pre-trained MultiMAE encoder with a linear classifier. Then, we fine-tune such a model following linear probing and end-to-end fine-tuning strategies as illustrated by Figure 4. During linear probing, the pre-trained encoder remains frozen, and only the parameters of the linear classifier are updated. In end-to-end fine-tuning, the pre-trained encoder and linear classifier parameters are updated. In the case of segmentation tasks, we plug a segmentation head into the pre-trained encoder. We perform fine-tuning, keeping the pre-trained encoder frozen (similar to linear probing) and standard end-to-end fine-tuning. The segmentation head consists of four ConvNeXt [11] blocks, which have demonstrated good alignment with ViT-based architectures [10].

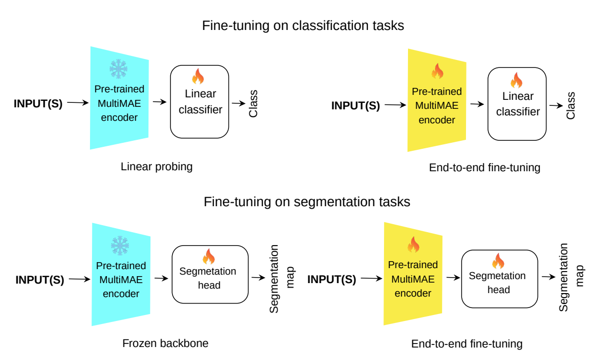


Fig. 4. **Fine-tuning setups for segmentation and classification EO tasks.** We follow standard end-to-end fine-tuning and linear probing for classification tasks. In segmentation tasks we perform fine-tuning keeping the pre-trained encoder frozen and end-to-end fine-tuning.

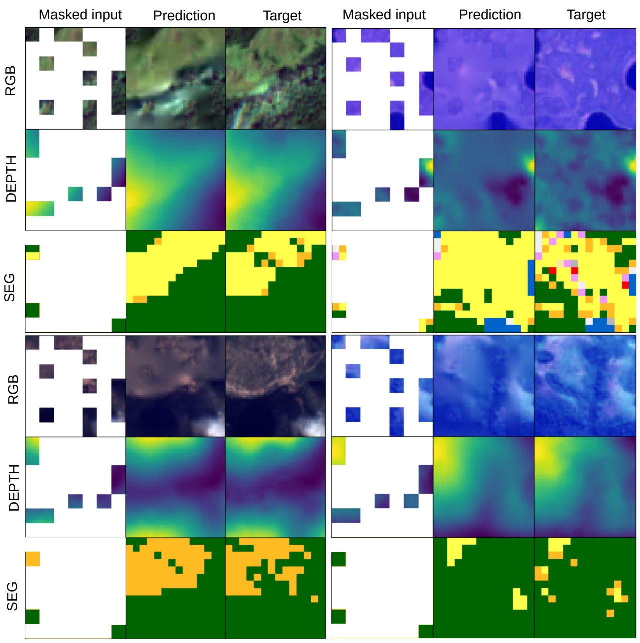


Fig. 5. **Visualisation of reconstructions across different input modalities.** Randomly chosen reconstructions of EO input modalities after pre-training MultiMAE. The first and fourth columns depicts the masked input for RGB, DEPTH, and SEG modalities. The second and fifth columns show the reconstructed image using our approach. The third and sixth columns display the corresponding ground truth (unmasked input).

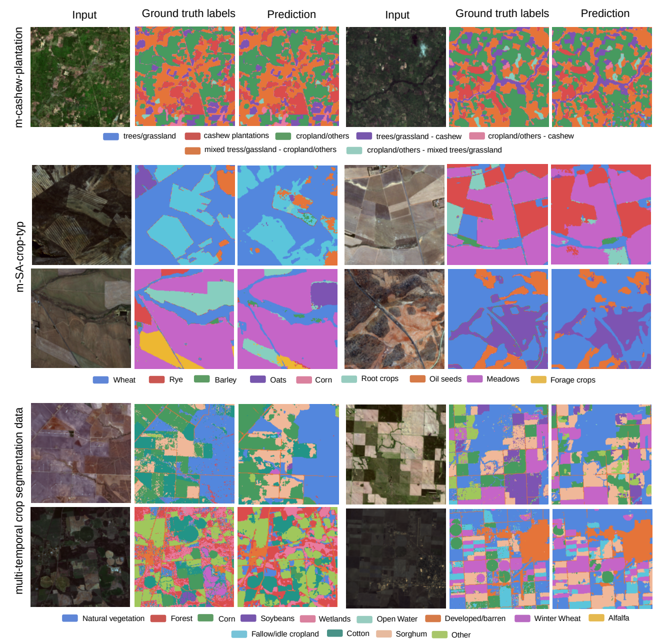


Fig. 6. **Visualisations for segmentation tasks.** The figure visualises the predictions after fine-tuning our approach with different segmentation datasets. The first column depicts an RGB representation of the input; the second column shows the ground truth segmentation labels from the respective dataset, and the third column depicts the predicted ones by our model. Each dataset group includes a legend showing the colour code for the labels used. Labels for m-cashew-plantation correspond to specific areas useful for tracking changes in land cover. In the case of the last two datasets, segmentation labels represent crop types mostly.

4. QUALITATIVE RESULTS

4.1. Pre-training visualisations

Figure 5 visualises randomly picked reconstructions produced by our approach. For simplicity, we only include reconstructions for RGB, DEPTH and SEG modalities within the figure. However, the pre-training stage involves the six modalities described in subsection 2.1. Note that these representations serve only illustrative purposes since they come from the training data. Based on visualisations from Figure 5, we can notice mostly accurate reconstructions across all input modalities, which is the intended goal of the self-supervised pre-training.

4.2. Qualitative results on segmentation tasks

We visualise some of the outputs after fine-tuning our approach for segmentation tasks. Figure 6 illustrates results for each of the three datasets that we used, namely m-cashew-plantation, m-SA-crop-type, and multi-temporal crop segmentation [12]. The first column on the figure depicts a representative RGB version of the inputs. However, note that for fine-tuning, as described in the main document, S2-derived modalities were used. Specifically, the input consists of RGB, IRED, SIRED, and EB (S2-derived) modalities for m-cashew-plantation and m-SA-crop-type datasets. For the multi-temporal crop segmentation dataset, input involves RGB, IRED, and DEPTH modalities (where depth corresponds to pseudo-labels).

5. REFERENCES

[1] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27811–27819.

[2] Qunming Wang, Wenzhong Shi, Zhongbin Li, and Pe-

ter M Atkinson, “Fusion of sentinel-2 images,” *Remote sensing of environment*, vol. 187, pp. 241–252, 2016.

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

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[5] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore, “Google earth engine: Planetary-scale geospatial analysis for everyone,” *Remote sensing of Environment*, vol. 202, pp. 18–27, 2017.

[6] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al., “Geo-bench: Toward foundation models for earth monitoring,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, “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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[8] 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*, vol. 35, pp. 197–211, 2022.

[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*, 2022, pp. 16000–16009.

[10] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir, “Multimae: Multi-modal multi-task masked autoencoders,” in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[11] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[12] Michael Cecil, Hanxi (Steve) Kordi, Fatemehand Li, Sam Khallaghi, and Hamed Alemohammad, “HLS Multi Temporal Crop Classification,” Aug. 2023.