How Effective is Pre-training of Large Masked Autoencoders for Downstream Earth Observation Tasks?

Jose Sosa[1]
jose.sosa@uni.lu
Mohamed Aloulou[1]
mohamed.aloulou@ext.uni.lu
Danila Rukhovich[1]
danila.rukhovich@uni.lu
Rim Sleimi[2]
rsleimi@hydrosat.com
Boonyarit Changaival[2]
bchangaival@hydrosat.com
Anis Kacem[1]
anis.kacem@uni.lu
Djamila Aouada[1]
djamila.aouada@uni.lu

[1] The Interdisciplinary Centre for Security, Reliability and Trust (SnT) University of Luxembourg Luxembourg

[2] Hydrosat Luxembourg

Abstract

Self-supervised pre-training has proven highly effective for many computer vision tasks, particularly when labelled data are scarce. In the context of Earth Observation (EO), foundation models and various other Vision Transformer (ViT)-based approaches have been successfully applied for transfer learning to downstream tasks. However, it remains unclear under which conditions pre-trained models offer significant advantages over training from scratch. In this study, we investigate the effectiveness of pre-training ViT-based Masked Autoencoders (MAE) for downstream EO tasks, focusing on reconstruction, segmentation, and classification. We consider two large ViT-based MAE pre-trained models: a foundation model (Prithvi) and SatMAE. We evaluate Prithvi on reconstruction and segmentation-based downstream tasks, and for SatMAE we assess its performance on a classification downstream task. Our findings suggest that pre-training is particularly beneficial when the fine-tuning task closely resembles the pre-training task, e.g. reconstruction. In contrast, for tasks such as segmentation or classification, training from scratch with specific hyperparameter adjustments proved to be equally or more effective.

1 Introduction

Self-supervised learning has been widely used for NLP [6, 13, 52] and subsequently for computer vision tasks [16, 23, 30, 38, 41, 48]. This practice aims to learn robust representations through pre-training models on large amounts of unlabelled data, followed by fine-tuning on particular downstream tasks where labels are available [25]. Initial successful attempts to apply the pre-training fine-tuning paradigm in the visual domain focused on various pretraining tasks [2, 22, 23]. However, thanks to the introduction of Vision Transformers (ViT) [15] and its posterior use for Masked Autoencoders (MAE) [23], the reconstruction task becomes a typical option for pre-training [16, 18, 47, 50].

Pre-training and subsequent fine-tuning of large ViT-based MAEs require significant computing resources [40]. Although this cost is justified for standard tasks like classification on datasets such as ImageNet [12], the benefits in other contexts, like the medical domain [31] and Earth Observation (EO) [11, 14], are less clear. In the case of EO, studies typically compare the performance of fine-tuning pre-trained ViT-based models against training from scratch well-known backbones such as ResNet [21], ConvViT [17], and U-Net [6]. However, these studies often lack rigorous justification for the performance gains resulting from pre-training, particularly missing basic hyperparameter tuning experiments [10, 28, 42]. This raises questions about the cost-effectiveness of pre-training for downstream tasks.

In this study, we investigate the effectiveness of relying on pre-trained large ViT-based MAEs for downstream tasks in the EO domain. Our objective is to determine whether training models from scratch for specific downstream tasks, with certain hyperparameter adjustments, can match or surpass the performance of initialising these from pre-trained large ViT-based MAEs. For this analysis, we focus on Prithvi [28], a foundation model that has been successfully applied to various segmentation and reconstruction tasks. Additionally, we analyse SatMAE [10], another large MAE ViT-based model designed for classification tasks. Although SatMAE is not considered as a foundation model, its structural similarity to Prithvi (both built on the ViT-based MAE architecture) makes it suitable for comparison in our study. We maintain the original distinction between the models, using Prithvi [28] for segmentation and reconstruction tasks, and SatMAE [10] for classification.

2 Related Work

Self-supervised pre-training of ViT-based models, particularly MAE, has proven beneficial in general settings [16, 23, 55, 58] relying on standard datasets such as ImageNet [12]. Consequently, this approach has been widely explored in other specific domains, including the medical field [31, 54] and, most relevant to our study, EO [10, 28, 42, 45, 53]. The vast amount of unlabelled data available for EO and remote sensing has enormously benefited the scaling of pre-training MAEs and other ViT-based models [29], encouraging the proliferation of many foundation models [3], such as Prithvi [28], SpectralGPT [26], S2MAE [37], and SkySense [20]. However, pre-training foundation models for EO requires large volumes of data and computing power [40], restricting their development to well-resourced research groups [33, 34]. Therefore, pre-traning of similar models but with less parameters and 'smaller' datasets have also been popular choice for transfer learning. Examples include SatMAE [10], ScaleMAE [45], Cross-Scale MAE [49], and SatMAE++ [42].

In parallel to the rise of foundation models for computer vision tasks [32, 36, 43, 44, 56, 60], many studies have also surged analysing their capabilities for transfer learning on

downstream general domain tasks [33, 57], as well as for specialised domains [27, 59]. For example, Huix *et al*. [27] evaluate the performance of popular foundation models such as SAM [32], SEEM [60], and DINO [43] on multiple medical datasets, finding that not all foundation models are suitable for transfer learning to downstream tasks in the medical domain. Similarly, Zhang *et al*. [59] conduct an in-depth analysis of the opportunities and challenges of using large pre-trained models in the medical field. More broadly, Chen *et al*. [7] and later Touvron *et al*. [51] introduce valuable studies related to pre-training and fine-tuning of ViTs. Their analyses offer insights into different combinations of hyperparameters that make the training of ViT-based models more stable and efficient.

Unlike the medical and general domains, unfortunately, there is still a lack of comprehensive studies analysing the pre-training of large models for EO [1, 39, 53]. One of the few such analyses is by Wang *et al*. [53], which evaluates the benefits of pre-training models with ImageNet for EO downstream tasks. Although this approach shares some similarities with ours, it differs in the nature of pre-training. Specifically, our study focuses on models that utilise domain-specific datasets during the pre-training stage, rather than relying on general-purpose datasets like ImageNet.

3 Proposed Study

Our study investigates the effectiveness of pre-training large ViT-based MAE models for downstream EO tasks. We analyse two settings, as illustrated in Figure 1. **Setting 1** involves initialising the encoder E with pre-trained weights obtained from a self-supervised pre-training stage. The encoder E is then coupled with a task-specific model M_i and fine-tuned using supervised learning. In **Setting 2**, the self-supervised pre-training stage is omitted, and E plus M_i are trained from scratch. We compare the corresponding task-specific metrics for both settings.

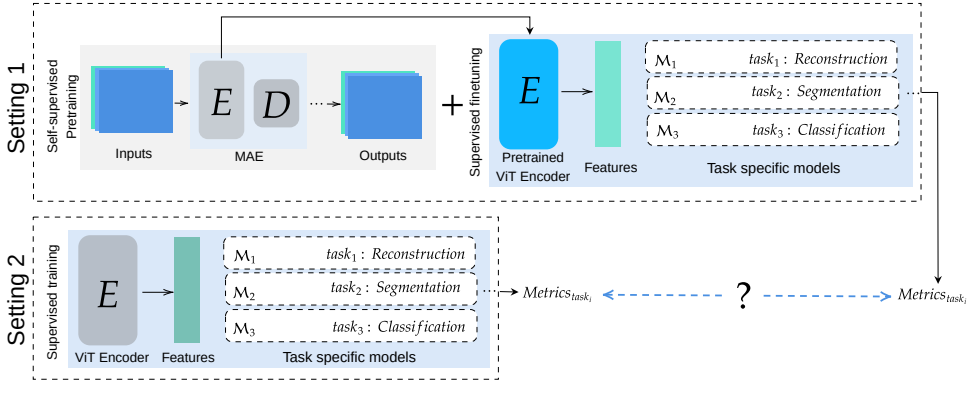

Figure 1: Settings for evaluation of ViT-based models. Setting 1 indicates that the encoder E coupled with M_i has been initialised with pre-trained weights. Setting 2 denotes the same, but without relying on pre-trained weights for E. Task related metrics have been compared for both settings to assess the effect of the self-supervised pre-training stage.

Due to the structural similarities between the models and the diverse datasets used for

their pre-training, we choose the encoders E from Prithvi [28] and SatMAE [10] for our experimental settings. Specifically, for Prithvi, we analyse its performance in the reconstruction and segmentation tasks outlined in [28], including temporal cloud gap imputation, temporal crop segmentation, flood mapping, and wildfire scar mapping. Additionally, we evaluate the robustness of the features learnt from cloud imputation fine-tuning when applied to crop segmentation. For SatMAE, we follow the original implementation and evaluate it exclusively on the classification task.

Note that for obtaining the results reported throughout section 4, all our experiments follow **Setting 2**. In other words, to build the models, we take the encoder E either from Prithvi or SatMAE, without initialisation, and couple it with a task-specific model M. Then, we perform supervised training of E and M several times with different sets of hyperparameters and report the corresponding metrics. In the case of **Setting 1**, since it represents standard pre-training and finetuning, we simply rely on metrics from the related original implementations. Experiments corresponding to **Setting 1** will be normally indicated with the name of the model used for initialising E (Prithvi or SatMAE), while results for **Setting 2** will be denote either as 'scratch' or 'scratch + hyp'.

Data and Considerations. We categorise our experiments into three main tasks: reconstruction, segmentation, and classification. Accordingly, we utilise different collections of data for each task. For reconstruction, we use data from the Multi-Temporal Cloud Gap Imputation dataset [19]. For segmentation, we utilise the Multi-Temporal Crop Segmentation dataset [35], Sen1Floods11 [4], and Wildfire Scar Mapping [46]. In the classification task, we rely on data from EuroSAT [24]. Given the high computational cost of experimenting with all possible combinations of hyperparameters for **Setting 2**, particularly when choosing ViTs for E, we strategically select key hyperparameters following some ideas from [7]. We focus on general hyperparameters like learning rate, learning rate scheduler, and batch size, as well as ViT-specific ones such as the number of heads and layers. To determine the optimal configuration, we experiment using small data subsets with various hyperparameter combinations selected for fixed ranges of values. Note that we consistently use a multistep learning rate (MultiStepLR) scheduler across all experiments, with decay occurring after 67% and 92% of the total number of epochs. A ViT-Large is also fixed as the backbone for most experiments, unless otherwise specified. For all other hyperparameter values we provide details in the proper following sections.

4 Results and Discussion

4.1 Multi-Temporal Cloud Gap Imputation

For the task cloud gap imputation, we rely on the approach proposed in [28]. In general, the task is simply image reconstruction, involving the use of a MAE to reconstruct regions covered by clouds in the given input image as illustrated in Figure 2. Note that in this case, the task for fine-tuning is exactly the same as that for pre-training. However, unlike the standard MAE masking [23], we follow [28] and use the binary cloud masks in the dataset [19] to build the masks needed for the inputs.

Following **Setting 2**, we train from scratch the components E and D (where D correspond to the task-specific model) of the model depicted in Figure 2. We use the same E and D architectures as in Prithvi and train these with data from the multi-temporal cloud imputation dataset [19]. We perform several experiments with different combinations of hyperparam-

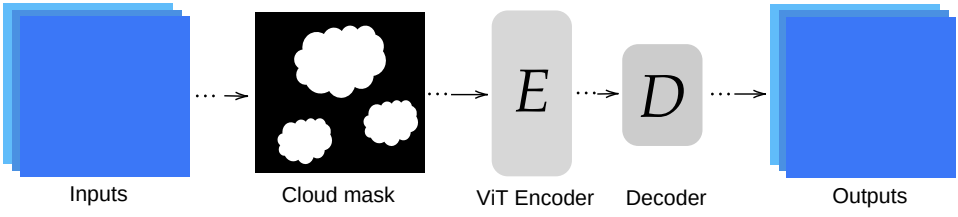

Figure 2: Training MAE for cloud imputation. Unlike the standard MAE which relies on fixed masking ratio, in this case we provide binary cloud masks for the inputs to train E and D from scratch.

eters. In summary, we lower the starting learning rate to 0.00005, replace the scheduler as mentioned in section 3, and reduce the number of heads and layers in the ViT encoder to 3 and 2, respectively. We train the model for 200 epochs, while maintaining the same batch size, patch size, and other hyperparameter values as indicated in [28]. We further investigate the impact of changing the encoder backbone for the MAE. In addition to the original ViT-Large backbone, we experiment with ViT-Base and ViT-Small [15]. Following [28], we evaluate all configurations using the test set from [19]. Table 1 reports the results in terms of mean absolute error (mae[1]) and structural similarity index (SSIM).

Initialisation	Backbone	mae	SSIM
Prithvi	ViT-Large	**0.020**	**0.972**
Scratch + hyp	ViT-Large	0.025	0.964
Scratch + hyp	ViT-Base	0.025	0.964
Scratch + hyp	ViT-Small	0.027	0.959

Table 1: Comparison of evaluation for cloud gap imputation. The first column indicates the initialisation used for E. The next columns provide details on the backbones and evaluation metrics.

Although altering the encoder backbone E results in a significant reduction in model parameters and accelerates the training time, it does not surpass the performance of the ViT-Large backbone used in [28]. We hypothesise that since the cloud imputation task is identical to the pre-training task, training from scratch with hyperparameter tuning has a minimal impact on the final performance. In this context, initialisation with pre-trained weights from Prithvi provides a beneficial effect on fine-tuning, compared to training from scratch.

4.2 Multi-Temporal Crop Segmentation

For the task of crop segmentation, we rely on the architecture depicted in Figure 3, which consists of a ViT-based encoder E coupled with a convolutional head. For training the model, we use labelled data from the multi-temporal crop segmentation dataset [65] as in [28].

In line with **Setting 2**, we train the model from scratch with the same hyperparameters as training with Prithvi initialisation (Scratch) and with some hyperparameters adjustments

[1]To avoid confusion with Masked Autoencoder (MAE), mean absolute error is denoted in lowercase.

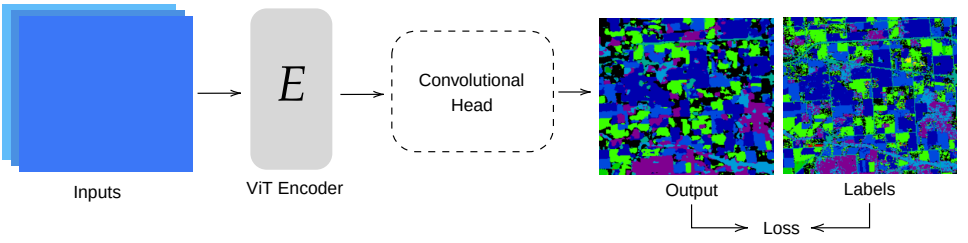

Figure 3: Standard emsemble for segmentation tasks. For crop classification, E has been coupled with a convolutional head, which could contain convolutional and linear layers. Labels are provided on the dataset for supervised training of the model.

(Scratch + hyp). In addition, we extend the analysis under **Setting 1** by exploring hyperparameters adjustments when initialising E with Prithvi (Prithvi + hyp). Table 2 summarises the hyperparameters values used for each of these experiments. After training, we perform evaluation using the corresponding data from [35]. Figure 4 shows the mean Intersection over Union (mIoU) for each of the 13 crop and land cover classes in the test set. The average values for all settings are indicated in the legend at the top of the figure.

Intialisation	Frames	Initial lr	Layers	Heads	Scheduler
Prithvi	3	$1.5e-5$	6	8	Polynomial
Prithvi + hyp	3	$1e-4$	6	12	MultiStepLR
Scratch	3	$1.5e-5$	6	8	Polynomial
Scratch + hyp	3	$1e-4$	6	12	MultiStepLR

Table 2: Summary of hyperparameters for different settings. The first column refers to the type of initialisation for the ViT encoder E, either from Prithvi [28] or from scratch.

According to Figure 4, the average mIoU for fine-tuning the model initialised with Prithvi's weights (Prithvi) is nearly the same as starting training from scratch (Scratch), with just a small difference of 0.6. Surprisingly, adjusting some hyperparameters in the latter setting (Scratch + hyp) leads to a significant increase in the average mIoU from 42.0 to 47.42. Additionally, using a combination of specific hyperparameters and initialisation from Prithvi (Prithvi + hyp) yields an average mIoU of 46.03, which is higher than the baseline performance (Prithvi), but still lower than the initialisation from scratch with the adjusted hyperparameters (Scratch + hyp).

Multi-temporal Crop Segmentation with Cloudy Data. The above experiments demonstrate that fine-tuning Prithvi can improve the performance of crop segmentation. Experiments reported in (subsection 4.1), showcase a better reconstruction of cloudy inputs with fine-tuning Prithvi on this task. However, the benefits of combining these tasks remain underexplored. In particular, it is unclear how effectively models pre-trained for cloud imputation perform on crop segmentation tasks when dealing with cloudy inputs (which is common in EO data).

To investigate this, we replicate the crop segmentation experiments described above with

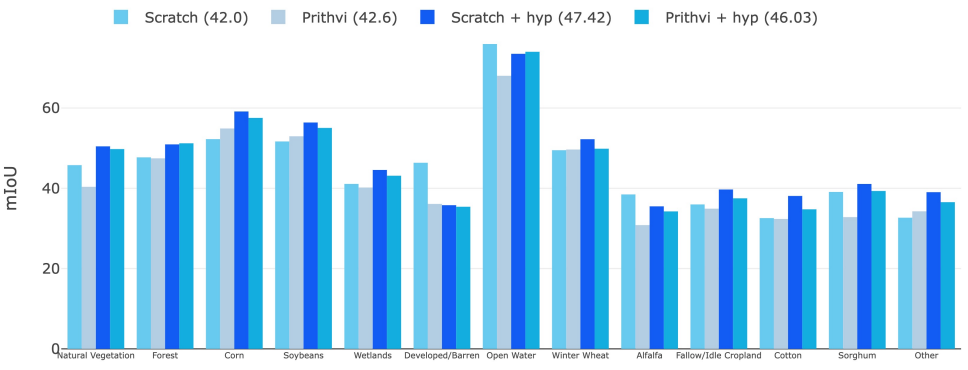

Figure 4: Comparison of performance of the model under different initialisations. We calculate the mIoU for each of the crop and land cover classes on the test set. Averages for each setting appear on the legend located at the top of the plot.

cloudy inputs. In particular, we simulate cloudy conditions using fixed masking ratios and apply these to crop segmentation data. We use masking levels of 30%, 60%, and 90% to retrain the crop segmentation model in Figure 3. For each masking ratio, we conduct experiments with different initialisation of the ViT encoder E: Prithvi + hyp, Scratch + hyp, and Prithvi + cloud finetuning. Note that the hyperparameters for the first two experiments are as specified in Table 2, while the last simply follows the training described above for standard crop segmentation.

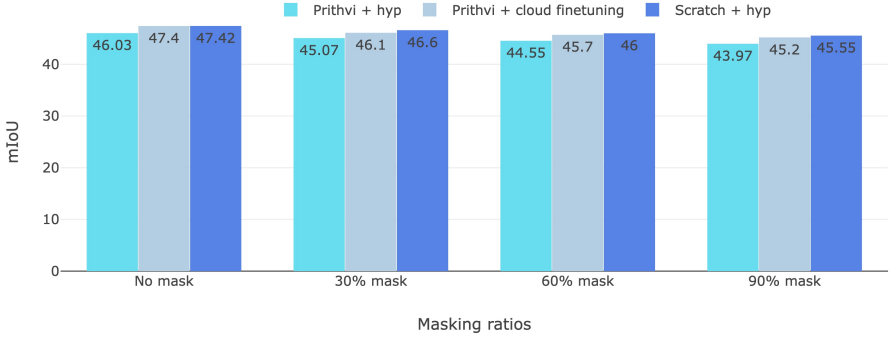

Figure 5: Different initialisation for crop segmentation with simulated cloudy data. Corresponding mIoU for each masking ratio with different initialisation for ViT Encoder.

As it can be observed from Figure 5, initialising E with pretrained Prithvi on cloud imputation (Prithvi + cloud finetuning) provides a slight improvement over standard Prithvi initialisation (Prithvi + hyp). However, contrary to what can be expected, training from scratch with hyperparameters adjustments (Scratch + hyp) is slightly better than other settings. Considering the extensive time and resources required for pre-training Prithvi, fine-tuning on the

cloud imputation task, and subsequently fine-tuning again for crop segmentation, it is clear that using pretrained Prithvi is not a cost-effective initialisation in this scenario.

4.3 Flood Mapping

We extend the segmentation experiments to flood segmentation using the dataset from [4], and following [28]. Unlike the data used for crop classification and cloud gap imputation, the Sen1Floods11 data set [4] lacks a temporal dimension, which requires the segmentation model to work with individual images. However, the model maintains the same general structure as the one used for crop segmentation, featuring a ViT encoder coupled with a few convolutional layers (Figure 3).

The approach imitates the one described in subsection 4.2, involving experiments under **Setting 2**, training the segmentation model from scratch plus some hyperparameters adjustments. Table 3 summarises the hyperparameters values used when training the model from scratch (Scratch + hyp). Note that in this case results when training from scratch (Scratch) with same hyperparameters as with Prithvi initialisation (Prithvi) are taken from the original paper.

Initialisation	Frames	Batch	Initial lr	Layers	Heads	Scheduler
Prithvi	1	4	$1.5e-5$	12	12	Polynomial
Scratch	1	4	$1.5e-5$	12	12	Polynomial
Scratch + hyp	1	8	$4e-5$	6	6	MultiStepLR

Table 3: Summary of hyperparameters for different experiments. First column denotes the type of initialisation for the ViT encoder E, either from Prithvi [28] or from scratch.

We train the model from scratch for up to 50, 100, and 400 epochs and evaluate each of them using the test set from [4]. We present results in Table 4, including all metrics and results reported in [28]. Similar to results with crop segmentation, few changes on the hyperparameters yield better performance than relying on Prithvi weights for initialisation. It is also worth noting that training from scratch significantly reduces the overall training time. Although this setting takes more epochs to match or surpass intialisation from pre-trained Prithvi in all the metrics, it is still more time efficient if we consider the fact that the Prithvi pre-training time is approximately 4.5 days [28].

Initialisation	Epochs	IoU (↑)	F1(↑)	mIoU(↑)	mF1(↑)	mAcc(↑)
Scratch	50	80.67	89.30	88.76	93.85	94.79
Prithvi	50	81.26	89.66	89.10	94.05	**95.07**
Scratch	500	82.97	90.69	90.14	94.66	94.82
Prithvi	500	82.99	90.71	90.16	94.68	94.60
Scratch + hyp	50	81.2	89.62	89.1	94.05	94.84
Scratch + hyp	100	82.15	90.26	89.73	94.42	94.93
Scratch + hyp	400	**83.11**	**90.78**	**90.24**	**94.72**	95.03

Table 4: Results for different initialisation of ViT encoder within segmentation model. Best results for each metric appear in bold.

4.4 Wildfire Scar Mapping

For wildfire scar segmentation experiments, we use data from the wildfire scar mapping dataset [26]. Following the same approach as for flood mapping, we train the model from scratch using the hyperparameters specified in the third row of Table 5.

Initialisation	Frames	Batch	Initial lr	Layers	Heads	Scheduler
Prithvi	1	4	$1.3e-5$	12	12	Polynomial
Scratch	1	4	$1.3e-5$	12	12	Polynomial
Scratch + hyp	1	8	$5e-5$	6	6	MultiStepLR

Table 5: Summary of hyperparameters for different experiments. First column refers to the type of initialisation for the ViT encoder E, either from Prithvi [28] or from scratch.

As shown in Table 6, training model from scratch for 100 epochs with some hyperpararemeters adjustments (Scratch + hyp) eventually outperforms all the metrics reported in [28] for wildfire scar mapping (Prithvi and Scratch). Although the model trained from scratch requires twice as many epochs to surpass its counterpart's performance, it is still convenient when considering the time required for pre-training Prithvi.

Initialisation	Epochs	IoU(↑)	F1(↑)	mIoU(↑)	mF1-score(↑)	mAcc(↑)
Prithvi	50	73.62	84.81	84.84	91.40	92.48
Scratch	50	72.26	83.89	84.01	90.87	92.41
Scratch + hyp	100	**73.99**	**85.05**	**85.41**	**91.72**	**93.79**

Table 6: Performance comparison of different model initialisations for wildfire scar mapping.

4.5 Land Cover Classification

In previous experiments we focus on reconstruction and segmentation tasks relying on Prithvi's encoder which is either initialised from scratch or pre-trained. Unlike typical pre-training approaches for large models, which usually rely on well-established datasets, the EO domain lacks a standardised dataset for pre-training. In the case of Prithvi, it has been pretrained with data from the NASA's HLS V2 L30 product [9]. To demonstrate that our findings are not specific to any pretrained dataset or finetuning task, we utilise SatMAE, a structurally similar large ViT-based MAE model to Prithvi, but pre-trained with different data. In particular, we use the SatMAE encoder, pre-trained on data from [8], for the land cover classification task. Following the same strategy as with segmentation experiments, we follow **Setting 2** to train from scratch a ViT model for classification with some hyperparameter adjustments (details of the model used could be found in [10]). Specifically, we modify the initial learning rate to $6e-4$ and keep the MultiStepLR scheduler. We train and test the model using the EuroSAT dataset [24]. We experiment with both RGB and multispectral data, and report the top-1 accuracy in Table 7, comparing the results with those of the model initialised with pre-trained ViT-encoder from SatMAE [10] (**Setting 1**).

Based on the results from Table 7, training the model from scratch with RGB data yields better performance when compared to initialisation from pre-trained SatMAE. Notably, initialisation from Scratch + hyp outperforms the SatMAE initialisation, even when both are

Initialisation	Input	Epochs	Top-1 Acc (↑)
SatMAE	RGB	50	95.74
Scratch + hyp	RGB	50	95.78
Scratch + hyp	RGB	100	**97.00**
SatMAE	Multi Spectral	50	**98.98**
Scratch + hyp	Multi Spectral	50	97.28
Scratch + hyp	Multi Spectral	100	98.44

Table 7: Results for different initialisation of ViT encoder used for classification. Best results for each setting appear in bold.

trained for the same number of epochs. Conversely, when using multispectral data, pre-training provides a slight improvement in performance.

4.6 Discussion

Based on the results from various EO segmentation and classification downstream tasks, we can observe that using large ViT-based MAE pre-trained models (**Setting 1**) does not consistently outperform models initialised from scratch (**Setting 2**). Our findings indicate that pre-training tends to improve performance for downstream tasks closely aligned with the pre-training task, such as the *Multi-Temporal Cloud Gap Imputation* task. However, for most segmentation tasks—including *Multi-Temporal Crop Segmentation, Multi-Temporal Crop Segmentation with cloudy data, Flood Mapping,* and *Wildfire Scar Mapping*—initialisation from scratch, together with hyperparameter tuning, can achieve comparable or even superior results. Similarly, for land cover classification, **Setting 2** is beneficial when using RGB inputs. However, when using multi-sprectral data, initialisation from pretrained SatMAE shows better performance.

5 Conclusion

In this paper, we analyse the effectiveness of pre-training large ViT-based MAE models for downstream EO tasks, with focus on one foundation model (Prithvi) and SatMAE. We experiment on reconstruction, segmentation, and classification EO tasks, demonstrating that relying on large ViT-based MAE pre-trained models as initialisation does not consistently outperform models initialised from scratch. Given that our experiments involve a diverse range of datasets on finetuning and pre-training stages, we hypothesise that the limitations observed in pre-training MAE ViT-based models might be more related to model design than to the data itself. This suggests that better strategies for pre-training foundation models and other MAE ViT-based models for EO could enhance the benefits of the fine-tuning process for downstream tasks. However, it is important to note that this study is relatively small in scope. Future research should extend these findings by incorporating additional datasets and models, particularly for classification tasks.

Acknowledgements. This work is supported by FNR HPC BRIDGES project under the reference HPC_BRIDGES/2022/17978225/AI4CC. The experiments were performed on the Luxembourg national supercomputer MeluXina. Thanks to LuxProvide teams for their support.

References

[1] Philipe Ambrozio Dias, Abhishek Potnis, Sreelekha Guggilam, Lexie Yang, Henry Medeiros, Dalton Lunga, et al. An agenda for multimodal foundation models for earth observation. Technical report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2023.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[9] Martin Claverie, Junchang Ju, Jeffrey G Masek, Jennifer L Dungan, Eric F Vermote, Jean-Claude Roger, Sergii V Skakun, and Christopher Justice. The harmonized landsat and sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219:145–161, 2018.

[10] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=WBhqzpF6KYH.

[11] Isaac Corley, Caleb Robinson, Rahul Dodhia, Juan M Lavista Ferres, and Peyman Najafirad. Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3162–3172, 2024.

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

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Nikolaos Dionelis, Casper Fibaek, Luke Camilleri, Andreas Luyts, Jente Bosmans, and Bertrand Le Saux. Evaluating and benchmarking foundation models for earth observation and geospatial ai. *arXiv preprint arXiv:2406.18295*, 2024.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[17] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022.

[18] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10406–10417, 2023.

[19] Denys Godwin, Hanxi (Steve) Li, and Hamed Alemohammad. Multi-Temporal Cloud Gap Imputation With HLS Data Across CONUS, 2024.

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

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[23] 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.

[24] 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*, 12(7):2217–2226, 2019.

[25] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[26] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[27] Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7634–7643, 2024.

[28] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence, 2023. URL https://arxiv.org/abs/2310.18660.

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[30] Ahmet Serdar Karadeniz, Dimitrios Mallis, Nesryne Mejri, Kseniya Cherenkova, Anis Kacem, and Djamila Aouada. Picasso: A feed-forward framework for parametric inference of cad sketches via rendering self-supervision. *arXiv preprint arXiv:2407.13394*, 2024.

[31] Muhammad Osama Khan and Yi Fang. Revisiting fine-tuning strategies for self-supervised medical imaging analysis. *arXiv preprint arXiv:2307.10915*, 2023.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[33] Adam Kolides, Alyna Nawaz, Anshu Rathor, Denzel Beeman, Muzammil Hashmi, Sana Fatima, David Berdik, Mahmoud Al-Ayyoub, and Yaser Jararweh. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, 126:102754, 2023.

[34] 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*, 36, 2024.

[35] Hanxi (Steve) Li, Sam Khallaghi, Michael Cecil, Fatemeh Kordi, Paolo Fraccaro, Hamed Alemohammad, and Rahul Ramachandran. HLS Multi Temporal Crop Classification Model, August 2023. URL https://huggingface.co/ibm-nasa-geospatial/Prithvi-100M-multi-temporal-crop-classification.

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[37] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24088–24097, 2024.

[38] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[39] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[40] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.

[41] Mohamed Adel Musallam, Vincent Gaudillière, and Djamila Aouada. Self-supervised learning for place representation generalization across appearance changes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7448–7458, 2024.

[42] 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*, pages 27811–27819, 2024.

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[44] 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.

[45] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[46] Sujit Roy, Christopher Phillips, Johannes Jakubik, Paolo Fraccaro, Kumar Ankur, Ryan Avery, Wei Ji, Bianca Zadrozny, and Rahul Ramachandran. Prithvi 100M burn scar, August 2023. URL https://huggingface.co/ibm-nasa-geospatial/Prithvi-100M-burn-scar.

[47] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5484–5494, 2023.

[48] Jose Sosa and David Hogg. Self-supervised 3d human pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4787–4796, 2023.

[49] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[51] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2022.

[54] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023.

[55] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021.

[56] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[57] Yang Yuan. On the power of foundation models. In *International Conference on Machine Learning*, pages 40519–40530. PMLR, 2023.

[58] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, 131(5):1141–1162, 2023.

[59] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, page 102996, 2023.

[60] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.