Context Autoencoders for Pretraining Foundation Models for Satellite Imagery Tasks

John Waithaka, Scovia Achan, Maurine Wanjiku, M. Cynthia A. Kamikazi [1]

Abstract

Self-supervised pre-training has been shown to improve the performance of vision models on downstream tasks, especially where labelled data is scarce. It is therefore particularly relevant in the earth observation and satellite imagery domain where, though there are massive satellite imagery datasets, there are few and small labelled datasets. We use context autoencoder (CAE), a masked image modelling self-supervised pretraining scheme, on satellite imagery. Our experiments show that CAE performs comparably to masked autoencoder (our baseline) on both the image reconstruction pretext task and land-use classification downstream tasks, while achieving slightly better performance on a flood mapping segmentation tasks. The code is available at `https://github.com/johnGachihi/satellite-cae`.

1 Introduction

Pre-training has become increasingly important in building generalist models that can transfer knowledge to various downstream tasks. These pre-trained models, however, not only generalise well but also achieve higher accuracy on downstream tasks, especially where labelled data is scarce.

In the field of Earth observation and satellite imagery, there is a scarcity of large labelled satellite image datasets at the scale of ImageNet. Therefore, tasks based on satellite imagery stand to benefit from pre-trained models. Fortunately, there are massive open unlabelled satellite image datasets. This, along with self-supervised pre-training strategies for vision tasks, has enabled the development of pre-trained models which have been shown to improve performance in earth observation tasks such as flood mapping and land-use classification [1, 2].

Masked Autoencoders (MAE) are effective self-supervised learners for developing pre-trained models [3]. They learn by reconstructing images from a few random patches, which encourages understanding local patterns in an image as well as its global context. Jakubik et al. [1] and Cong et al. [2] use MAE on satellite image data and show superior performance in downstream earth observation tasks over models trained from scratch.

The context autoencoder (CAE), an extension on MAE, has been shown to perform better on downstream tasks compared to other self-supervised pre-training strategies, including MAE [4]. CAE adds a 'regressor' component between its encoder and decoder, which encourages the responsibility of representation learning to be on the encoder alone [4].

Despite superior performance over MAE on various downstream tasks outside the satellite imagery domain, CAE has not been used on satellite image data and tasks. Therefore, we compare the effectiveness of MAE and CAE pre-training strategies on satellite imagery by comparing their transfer performance on two downstream tasks - flood mapping and land-use classification.

The data used for pre-training and evaluation on the downstream tasks are images of the earth captured from the Sentinel-2 satellite platforms. The images contain reflectance values for 13 spectral bands including near-infrared, shortwave-infrared, red, green and blue.

Preprint. Under review.

2 Literature Review

2.1 Self-Supervised Pre-training

Pre-training is a process where a model is first trained on a large dataset to learn generalisable semantic features before being applied to a specific downstream task. Pre-training gives a model a head start by positioning it in a parameter space that biases it toward finding good minima on downstream tasks [5]. Importantly, in cases where the downstream task's training set is small and deep models tend to overfit, pre-training leads to minima with more generalisable performance [5].

Self-supervised pre-training is a variant of pretraining in which a pretext task defines a method to extract target labels directly from the data and a false task for which the model should be optimised. Notably, self-supervised pretraining does not require labelled data and is therefore appropriate in cases like ours where there is little labelled data but massive unlabelled datasets.

Many self-supervised pre-training strategies for computer vision have been proposed. These include deep clustering methods [6, 7] which involve learning representations as part of an EM clustering algorithm, spatial context prediction [8] involving the prediction, for example, of the spatial position of an image patch given another patch of the same image, transform-based methods that involve the prediction of applied transformations [9] or the original image given the transformed ones [10, 11], and the recently popular contrastive methods [12, 13] involving learning representations that maximise the similarity between augmented versions of the same image and disimilarity between different images. He et. al [13] show that MoCo, a variant of the contrastive methods, outperformed all prior self-supervised methods evaluated on fine-tuning accuracy on ImageNet-1K after pre-training on ImageNet-1M.

Of more relevance to our work are masked image modelling methods.

2.2 Masked Image Modelling

Masked Image modelling (MIM) methods, motivated by BERT [14] in masked language modelling, learn representations from images distorted by masking. Recent MIM methods are based on Transformers. Dosovitskiy et al. [15], in the ViT paper, and Bao et al. [16] with BEiT, explore masked patch prediction for self-supervised pre-training. BEiT outperforms previous self-supervised pre-training methods, including MoCo [16], evaluated on fine-tuning accuracy on ImageNet-1K.

2.3 Masked Autoencoder

The masked autoencoder [3] a recent variation of MIM, is a ViT-based autoencoder that reconstructs the original image given a masked version. First, as in standard ViT [15], an image is split into regular non-overlapping patches. Then a set of patches is randomly sampled (without replacement) following a uniform distribution, and the remaining patches are masked. The ratio of masked patches is high (e.g. 75%). The patches are embedded as in standard ViT with the addition of a positional embedding. The visible patch embeddings are processed by the MAE encoder's Transformer blocks. Then the visible patch embeddings outputted by the encoder along with the masked patch embeddings are passed through the MAE decoder Transformer blocks to predict the pixel values of the masked patches.

He et al. [3] show that MAE's performance is much better than MoCo and slightly better than BEiT on various downstream tasks, including semantic segmentation and classification.

2.4 Masked Autoencoders on Satellite Images

Several self-supervised pretraining strategies have been used on satellite imagery in prior work (e.g. [17, 18, 19].) We focus on MAE, making the assumption that since it outperforms the others on the ImageNet datasets it will outperform them on satellite imagery datasets too.

MAE has been used for satellite imagery in various ways. For example, Li et al. [20] build a foundation model for Synthetic Aperture Radar (SAR) automatic target recognition by using MAE on SAR data, Tseng et al. [21] build a lightweight foundation model pre-trained on optical, SAR, NDVI, climate reanalysis, land cover and topography time series data, and Reed et al. [22] add to standard MAE positional encodings and decoder to encourage scale-invariant representations.

SatMAE [23], our baseline, uses MAE on multispectral Sentinel-2 satellite imagery with the optional addition of a temporal dimension (i.e., a sequence of images). Compared to other pre-training strategies and models, SatMAE performs significantly better on downstream classification tasks and competitively on downstream segmentation tasks. In the classification task on fMoW-Sentinel, SatMAE achieves a top-5 accuracy of 85.17% where the second-best performing pretraining strategy (MoCo-v3 [24]) achieves 76.35%.

We select SatMAE as a baseline due to the accessibility of its pre-training dataset, which facilitates reproducibility.

2.5 Context Autoencoders

Like MAE, Context Autoencoders learn by reconstructing masked images using an encoder that takes in visible patches and outputs their representation, and a decoder that outputs the reconstructed masked patches. However, CAE adds a "regressor" between the encoder and decoder. This regressor takes the encoder's output as input and uses it to predict the representations of the masked patches. These predictions are then fed as input to the decoder, which uses them to reconstruct the masked patches. The additional component in CAE— prediction of the masked patches' representations within the representation space of the encoder output, rather than reconstruction directly from the encoder output as in MAE—encourages higher semantic understanding and also higher quality representations from the encoder by encouraging the responsibility of representation learning to be solely on the encoder.

Chen et al. [4] show CAE performs better than MAE on a variety of downstream tasks. For example, in segmentation on the ADE20K dataset, CAE achieves a mIoU of 54.7% while MAE achieves 53.6%. In classification on the Clipart dataset, CAE achieves an accuracy of 81. 84%, whereas MAE achieves 80.63%. We therefore seek to use CAE in the context of earth observation and compare its performance to MAE on segmentation tasks, specifically flood mapping and land-use classification.

3 Dataset

3.1 Dataset for Pretraining and Land Use Classification

The fMoW Sentinel dataset was adopted for pre-training as in SatMAE. This dataset comprises Sentinel-2 images with all 13 spectral bands and has 712,874 training images, 84,939 validation images and 84,966 test images. The dataset for pretraining was uniformly sampled across various geolocations from the full FMOW dataset, resulting in a subset of 380,606 images. This subset was used to pretrain both the MAE and CAE models, ensuring diverse geographic representation in the training data and validation was done on 84,967 images. The fMoW dataset comprises 62 scene categories/classes, including airport, crop field, golf course, and zoo.

For land use classification, we utilized a subset consisting of 59,401 images for training, 15,678 images for validation and 84,967 images as the test-set, ensuring a robust evaluation of model performance.

3.2 Datasets for Flood Mapping Downstream Task

The Sen1Floods11 dataset [25] was employed for the downstream flood mapping segmentation task. This dataset includes image chips from Sentinel-1 and Sentinel-2, along with binary pixel-level flood maps spanning 11 flood events across 14 biomes, 357 ecoregions, and 6 continents. For this study, we utilized 252 images for training, 89 images for validation, and 90 images for testing.

4 Evaluation metrics

To evaluate the performance of our pretrained models (MAE and CAE), we use the Mean Squared Error (MSE) during pretraining and accuracy or IoU (Intersection over Union) for downstream tasks.

4.1 Pre-training

Mean Squared Error(MSE): Employed to measure the average squared difference between the predicted pixel values and the ground truth. We used it in reconstruction tasks to evaluate how accurately the model recreates the input image from its latent representations. For reconstruction quality:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where N: Total number of patches, y_i: Ground truth value of patch i, $\hat{y}_i$: Predicted value of patch i

4.2 Fine-tuning

Accuracy measures the proportion of correct predictions among all predictions made by the model. This metric was used to evaluate the model's performance on land-use classification task:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where **TP** indicates correctly predicted flooded areas, **TN** non-flooded areas, **FP** false alarms, and **FN** missed flooded areas.

Intersection over Union (IoU), also known as the Jaccard index, is our metric for evaluating flood segmentation task performance. IoU is widely used in computer vision tasks like segmentation, object detection, and tracking because of its ability to directly quantify the overlap between predicted and ground truth regions

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

where:
Area of Intersection: Pixels that are labeled as flooded in both the predicted mask and ground truth.
Area of Union: Total pixels labeled as flooded in either the predicted mask or the ground truth (includes true positives, false positives, and false negatives).

5 Baseline Model

5.1 Architecture

The framework adopted the SatMAE framework which is based on MAE to pre-train multi-spectral satellite imagery [26]. The approach shows improved performances for self-supervised learning by up to 7% in comparison to the current state of art for existing benchmark datasets. Additionally, there was 14% improvement for downstream remote sensing tasks such as land cover classification using transfer learning. Unlike the MAE which usually processses RGB images, the SatMAE can process images with multiple spectral bands like satellite images [3]. For instance the Sentinel-2 images has 13 bands with varying spatial resolutions and wavelengths.

To fully capture the information from each of the spectral bands, we used group channels which are created by organizing the spectral bands into subsets. Each subset is processed to create a sequence of embedded tokens. The subset tokens are then concatenated to create a final set of tokens for the group channels which are used for spectral encoding. Compared to MAE, SatMAE does positional encoding on each group channel by concatenating the spectral encoding information to the $x_{k,i}$, $y_{k,i}$ positions to obtain the final dimensional data which is then masked. The masking strategies used for this experiment are consistent masking and independent masking. In consistent masking, each image is masked separately with the regions masked being consistent across all the images while for independent masking the regions masked differ across every image. The Figure 1 shows the MAE layout as adopted in SatMAE.

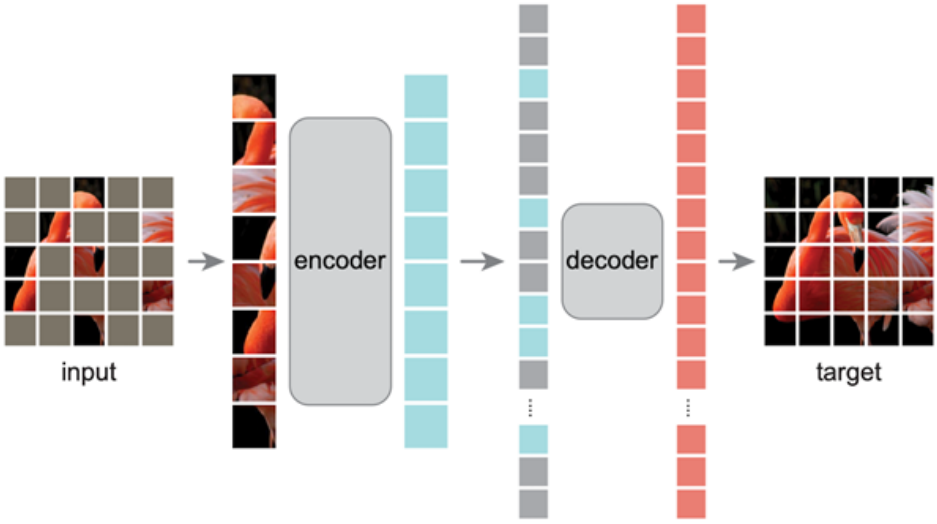


Figure 1: Masked Autoencoder (MAE) architecture. [26]

Adopting ViT transformer, 75% of the image is masked randomly following a uniform distribution. The encoder maps the visible patches to their corresponding latent representations. The inputs to the decoder are the latent representations of the visible patches and the mask tokens which are used for the image reconstruction. The architecture of the decoder is flexible and independent to the encoder as it's only pretrained to do reconstruction. They are smallers in size in comparison to the encoder to reduce pretraining time on reconstruction.

5.2 Loss Function

The loss function used for the baseline (MAE) is the mean squared error (MSE). This measures the reconstruction accuracy by computing the squared difference between the ground truth image I and the decoder's reconstructed output $\hat{I}$, focusing only on the masked patches:

$$MSE = \frac{1}{N} \sum_{i \in M} (I_i - \hat{I}_i)^2 \tag{1}$$

where:

- I represents the original input image,
- $\hat{I}$ is the reconstructed image produced by the decoder,
- M denotes the set of masked patches,
- N is the total number of pixels within the masked patches, and
- i indexes the pixels in the masked patches.

Unlike traditional approaches that evaluate loss across the entire image, the MSE here is computed exclusively on the masked regions ($i \in M$). The calculation operates on a per-pixel basis rather than per-patch, allowing for finer-grained reconstruction evaluation. Furthermore, the normalization factor N ensures that the loss is appropriately scaled, regardless of the number or size of masked patches.

5.3 Baseline Implementation

Following the setup detailed in the SatMAE paper, we implemented the MAE-based baseline model for pretraining on multi-spectral Sentinel-2 imagery, specifically utilizing the FMOW-Sentinel dataset. Our implementation does not necessarily use the original data size and splits for training, validation,

and testing as detailed in dataset section, but it has followed the same pretraining pipeline, including a masking ratio of 75% and reconstruction of masked patches using MSE loss.

Our baseline achieved an accuracy of 76.86% on the FMOW-Sentinel classification task after 30 epochs of fine-tuning, which is much less than the results reported in the original SatMAE paper. This discrepancy reflects the inherent challenges of reproducing results due to variations in computational resources which enforced the adaptation of smaller dataset size in pretraining and other computational resources dependent parameters.

6 Proposed Model

6.1 Architecture

The proposed model for our task is the context autoencoder (CAE). The CAE works in two parts namely; learning the encoder and completing pretraining tasks [4]. Similar to the MAE, the CAE takes in a masked image which is fed to an encoder and images are reconstructed using a decoder. However, the main difference is that the CAE has an addition of a regressor which predicts the masked patches which makes the architecture an encoder-regressor-decoder as shown in Figure 2.

The **encoder** maps the visible part of the satellite images X_v to the latent representation Z_v. We adopt the vision transformer (ViT) for the encoder, similar to what was used in the original research [4]. The ViT initially embeds the visible parts of the image by linear projection as patch embeddings and then adds positional embeddings P_v. These combined embeddings are sent into a sequence of transformer blocks which are based on self-attention to generate Z_v.

The **regressor** predicts latent representations Z_m for masked patches using visible patch representations Z_v from the encoder, conditioned on masked patch positions. The decoder is formed from transformer blocks based on cross-attention, with learned mask tokens Q_m serving as initial queries for the masked patches. The keys and values are derived from Z_v and previous mask queries. Positional embeddings of masked patches help compute cross-attention weights, while Z_v remains unchanged throughout.

The **decoder** maps the latent representations Z_m of masked patches to predicted masked patches Y_m. Like the encoder, it consists of transformer blocks based on self-attention, followed by a linear layer for target prediction. The decoder uses only the latent representations of masked patches (from the latent contextual regressor) and the positional embeddings of these patches, without directly incorporating information from the visible patches.

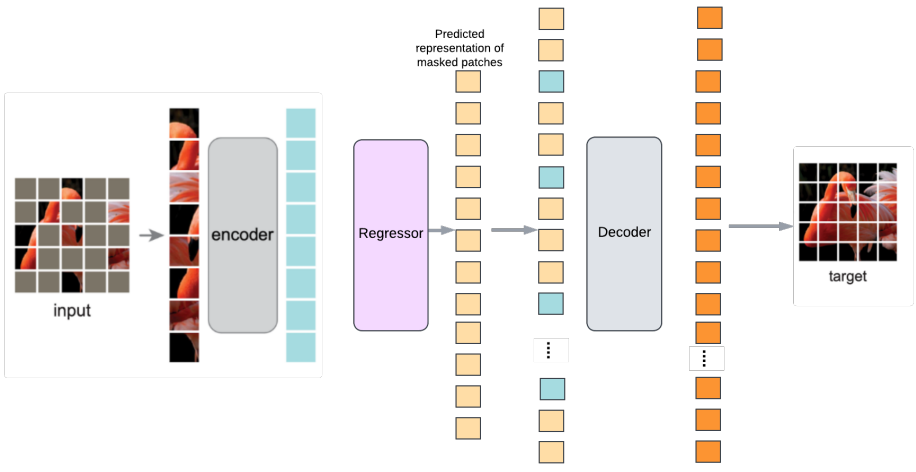


Figure 2: Context autoencoder (CAE) architecture [4]

6.2 Loss Function

The loss function employed for the CAE is the same as that defined for the MAE in Equation 1. Evaluates reconstruction accuracy by calculating the square difference between the ground truth image I and the reconstructed output $\hat{I}$, considering only masked patches.

6.3 Pre-training

The pre-training process utilized the Functional Map of the World (FMoW) dataset, consisting of 380,606 Sentinel-2 13-band images uniformly sampled across geolocations from the original dataset used in [4]. While the original CAE was designed for RGB images, we adapted it to accommodate the additional spectral channels present in satellite imagery.

A ViT-Base backbone with a patch size of $P = 16$ was used, resizing all images to 224×224 pixels. The model was trained for 100 epochs with a masking ratio of 0.75, aligning with what was used in the original CAE method. The pre-training objective was to reconstruct the masked patches, with the loss calculated exclusively on these regions using the mean squared error (MSE). The outcomes of pre-training are discussed in the results section.

6.4 Fine-tuning

The fine-tuning process involved two downstream tasks: flood mapping and land-use classification.

For land-use classification, we used the FMoW-Sentinel dataset, described in the dataset section. The model had a top-5% accuracy of 73.283% for, higher than 73.121% for MAE and 56.101% when trained from scratch (without fine-tuning).

The loss function employed for land-use classification was the **cross-entropy loss**. The formula for the cross-entropy loss adopted from PyTorch [27]:

$$\text{CrossEntropyLoss} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(x_i)}{\sum_{j=1}^{C}\exp(x_j)}\right)$$

where:

- x_i is the raw output (logits) for class i.
- y_i is the true class index.
- N is the number of samples in the batch.
- C is the total number of classes (62 classes in the dataset).

For flood mapping, we used the **Sen1Floods11** dataset, which includes Sentinel-2 13-band images paired with corresponding flood map labels. To address the class imbalance between 'flood' and 'no-flood' regions, we employed the **weighted cross-entropy loss**. The formula for this loss function is:

$$\text{CrossEntropyLoss} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log\left(\frac{\exp(x_i)}{\sum_{j=1}^{C}\exp(x_j)}\right)$$

where:

- C is the total number of classes, where $C = \{\text{'flood', 'no-flood'}\}$.
- w_{y_i} is the class-specific weight, with w values set to $w = [0.7, 0.3]$.

By assigning higher weights to the less frequent class ('flood'), this approach ensures that these regions have a larger impact on the loss calculation, effectively mitigating class imbalances during model training.

7 Results

7.1 Pre-training

Fig. 3 shows the train loss graph by epoch for both CAE and MAE, and Table 1 shows the reconstruction loss (also the metric) of two schemes on the test set. We see that the reconstruction performances are almost equal.

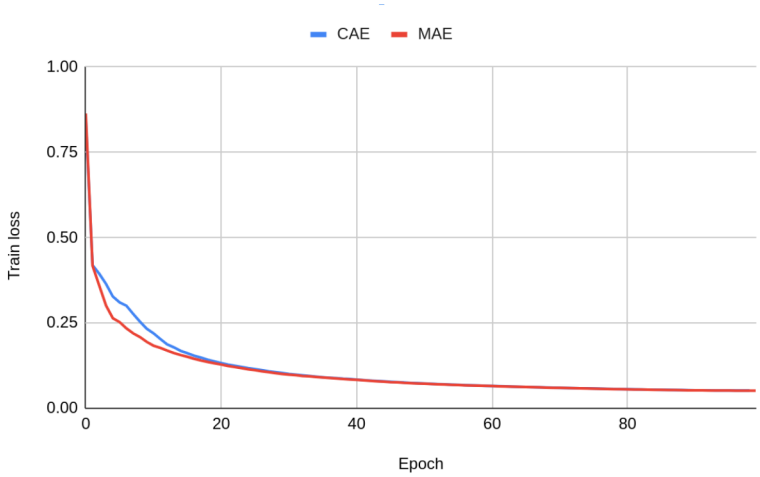
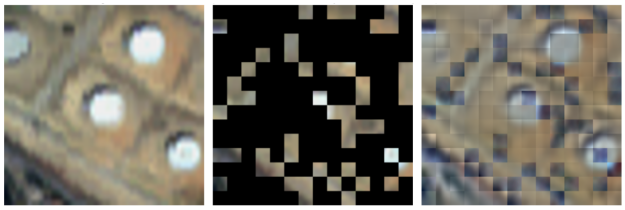


Figure 3: CAE and MAE train loss by epoch

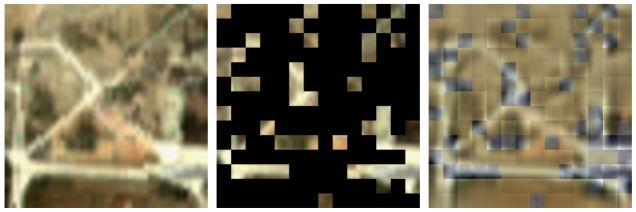
Table 1: Reconstruction error on test set

Model	Mean square error
CAE	1.0452
MAE	1.0426

Fig. 4 shows examples of reconstruction for both CAE and MAE.



(a) CAE reconstruction example



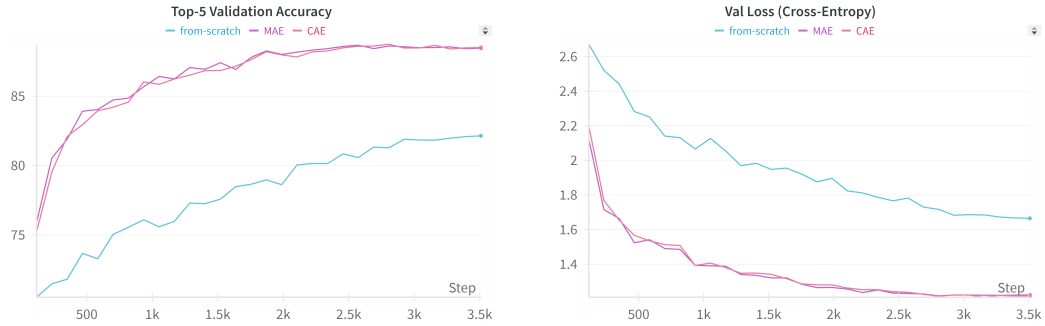
(b) MAE reconstruction example

Figure 4: Reconstruction examples for both CAE and MAE. We show the original image (left), the masked image (middle) and the reconstructed image (right).

7.2 Fine-Tuning

7.2.1 Land-Use Classification on fMoW-Sentinel

Fig. 5 shows the validation accuracy and loss graphs. We see that, whereas CAE and MAE have similar performance, the from-scratch case performance is much worse. In Fig. 5a we see that both CAE and MAE have a higher start, a higher slope, and converge at a higher accuracy than the from-scratch model, showing that all the benefits of pre-training are achieved.



(a) Top-5 validation accuracy

(b) Validation loss (cross-entropy)

Figure 5: Validation accuracy and loss

Table 2 shows the top-1 and top-5 accuracy for the three cases on the test set. We see that CAE and MAE have similar performance, with MAE outperforming CAE on top-1 accuracy by 0.05%.

Table 2: Top-1 and top-5 accuracy on the test set

Model	Top-1 Accuracy	Top-5 Accuracy
CAE	46.168	73.283
MAE	46.210	73.121
from-scratch	28.504	56.101

Fig. 6 shows the train loss graph for CAE, MAE and the from-scratch case.

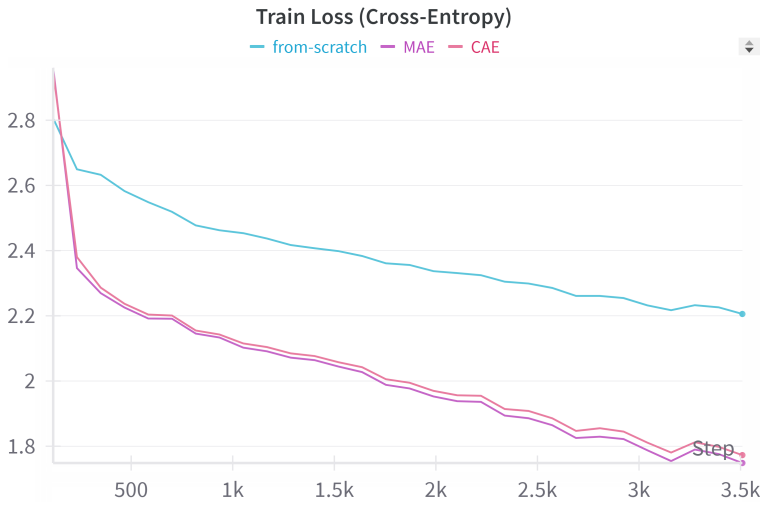


Figure 6: Train loss for classification on fMoW-Sentinel

7.2.2 Segmentation on Sen1Floods11

Table 3 shows the performance of CAE, MAE and the from-scratch cases on the Sen1Floods11 test set. For each case, we saved the model at each epoch during training and selected the one with the best performance on the validation set. We see that CAE outperforms MAE by 0.82%. We also see that at 50 epochs, the from-scratch case outperforms MAE.

Table 3: mIoU on the Sen1Floods11 test set

Model	35 epochs	50 epochs
CAE	77.68	77.68
MAE	76.86	76.86
From scratch	76.7	77.64

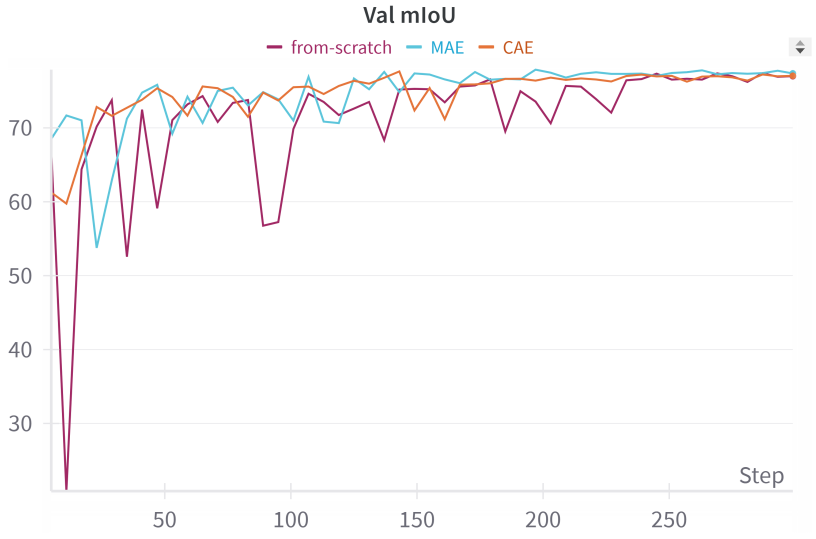
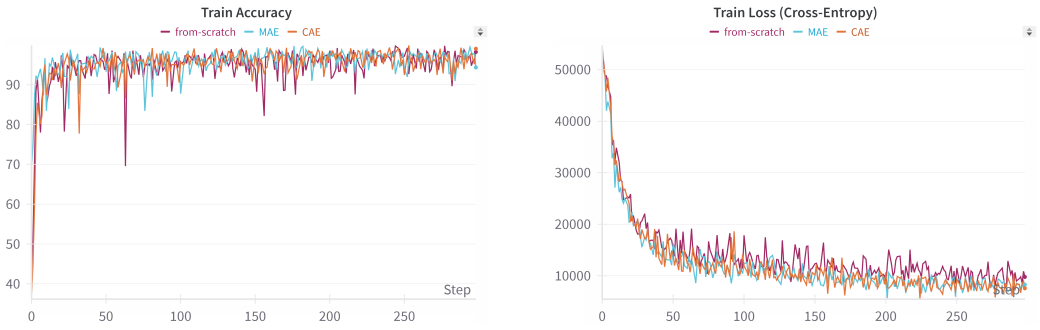


Figure 7: Validation mIoU on Sen1Floods11

Fig. 8 shows the train graphs for fine-tuning on Sen1Floods11.



(a) Train accuracy

(b) Test loss

Figure 8: Train accuracy and cross-entropy loss on Sen1Floods11

8 Discussion

Our experiments provide some insight into how CAE compares with MAE for satellite imagery tasks. The pretext task results show similar reconstruction capabilities between CAE (MSE: 1.0452) and MAE (MSE: 1.0426). This comparable performance is noteworthy. If the author's claims are true—that CAE's encoder, unlike MAE's, is solely responsible for representation learning—then CAE's encoder should be more powerful than MAE's on downstream tasks.

In both downstream tasks, the architectures exhibited similar performance. CAE achieved marginally lower accuracy in classification (46.168% vs 46.210%) and marginally higher mIoU in segmentation (77.68% vs 76.86%). However, these modest improvements must be weighed against CAE's higher computational cost (for pre-training) introduced by its regressor.

Some factors that may have influenced our results are that (1) we did not implement CAE's alignment mechanism, and (2) we used MAE's masking strategy on CAE rather than CAE's native approach. These factors, driven by time and resource constraints, probably put CAE at a disadvantage in our experiments and should be implemented in future.

Lastly, ablations, especially on the downstream tasks, were limited by time and resources. Therefore, the results presented should be viewed as preliminary rather than definitive.

9 Future Work

A significant limitation of our project was the constraints on resources and time, which restricted the scope of our experiments. For future work, pretraining the two models for more epochs and thoroughly evaluating their performance could provide more robust evidence to support the claim that CAE outperforms MAE in multispectral satellite imagery tasks. Additionally, future efforts could include implementing the alignment task in CAE, as proposed by [4]. This task aims to enhance the performance of the encoder model, particularly for multispectral satellite imagery tasks, and could further improve the effectiveness of CAE in this domain. Furthermore, we conducted only a limited number of downstream tasks, focusing on land-use classification and flood segmentation. Expanding the application of the pretrained models to additional satellite imagery downstream tasks represents another promising direction for future work. Finally, pretraining the MAE and CAE models on domain-specific data is anticipated to produce superior and more comprehensive results for segmentation tasks compared to training models from scratch. This advantage arises from the ability of pretrained models to adapt more effectively to the unique characteristics of the data, including differences in input image resolution, thereby enhancing overall performance. .

10 Conclusion

The objective of this project was to compare the performance of CAE and MAE on satellite imagery tasks. Our experiments show that CAE and MAE exhibit comparable performance when pretrained for 100 epochs. Both models employed two pretraining subtasks: generating representations using the encoder and reconstructing images using the decoder. As noted in future work, extending the training to more epochs is expected to improve the performance of both models for pretraining and downstream tasks. Additionally, the results demonstrate that pretrained models consistently outperform models trained from scratch, emphasizing the benefits of pretraining. These advantages include better initial performance (higher start), faster improvement during training (higher slope), and superior final performance (higher asymptote). .

Notes

[1]Scovia Achan reviewed the CAE literature and contributed to its design and implementation. Maurine Wanjiku configured and fine-tuned the baseline model (MAE). John Waithaka set up and ran the baseline model's pretraining experiments. M. Cynthia A. evaluated the CAE's performance on classification and segmentation tasks

References

[1] 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. 10 2023.

[2] 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. *Advances in Neural Information Processing Systems*, 35, 7 2022.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:15979–15988, 11 2021.

[4] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context Autoencoder for Self-Supervised Representation Learning. *International Journal of Computer Vision*, 132(1):208–223, 2 2022.

[5] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why Does Unsupervised Pre-training Help Deep Learning?, 3 2010.

[6] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *33rd International Conference on Machine Learning, ICML 2016*, volume 1, 2016.

[7] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, 2016.

[8] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised Visual Representation Learning by Context Prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[10] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9), 2016.

[11] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS, 2016.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2006.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. In *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[16] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEIT: BERT PRE-TRAINING OF IMAGE TRANSFORMERS. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[17] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-Aware Self-Supervised Learning. 2022.

[18] Oscar Mañas, 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 International Conference on Computer Vision*, 2021.

[19] Wenyuan Li, Hao Chen, and Zhenwei Shi. Semantic Segmentation of Remote Sensing Images with Self-Supervised Multitask Representation Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 2021.

[20] Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. SARATR-X: A Foundation Model for Synthetic Aperture Radar Images Target Recognition. 5 2024.

[21] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries.

[22] 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 International Conference on Computer Vision*, 2023.

[23] Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. October 2022.

[24] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9620–9629, 4 2021.

[25] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2020-June:835–845, 6 2020.

[26] 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, 2023.

[27] PyTorch Contributors. torch.nn.crossentropyloss, 2024. Accessed: 2024-12-10.