SELF-SUPERVISED REPRESENTATION LEARNING ON REMOTE SENSING PIXEL TIME SERIES WITH PATCH-BASED MASKING

Clinton Oduor*
Amini.ai
clinton@amini.ai

Jackline Tum*
Amini.ai
jackie@amini.ai

Collins Asega
Amini.ai
collins@amini.ai

Fancy Chepkoech
Amini.ai
fancy@amini.ai

ABSTRACT

We build on the patch-segmentation and channel-independence concepts introduced by PatchTST (Nie et al., 2023), adapting them to remote-sensing pixel time series for self-supervised representation learning. In our approach, each spectral band/index (e.g., Red, NDWI) is treated as a univariate sequence, split into patch tokens, then processed by a 600k-parameter transformer encoder. We randomly mask a substantial fraction of the patch tokens and train the transformer model to reconstruct the missing segments, capturing both short sub-seasonal dynamics within each patch and multi-seasonal context across the entire time series. Empirical results show that the learned representations transfer effectively to downstream tasks such as land-cover classification. We also discuss future directions, including foundation-scale training and integration of additional sensor modalities.

1 INTRODUCTION

Self-supervised learning has sparked breakthroughs across various modalities, including remote sensing (Szwarcman et al., 2025; Tseng et al., 2024; Lu et al., 2024; Huo et al., 2025). In this work, we adapt patch-based masking techniques (Nie et al., 2023) to remote sensing pixel time series, enabling the model to learn meaningful representations without labeled data. Our methodology leverages the temporal structure of remote sensing data, treating each spectral band as an independent time series and using a Transformer-based architecture to capture both local and global temporal patterns.

2 METHOD OVERVIEW

2.1 INPUT REPRESENTATION AND PATCH TOKENIZATION

We consider a pixel-level multivariate time series composed of M spectral bands or indices. Each band $m \in \{1, \ldots, M\}$ is treated as a univariate time series $\{x_1^{(m)}, x_2^{(m)}, \ldots, x_L^{(m)}\}$, where L denotes the temporal length (e.g., biweekly composites spanning multiple months). Bands include both raw reflectance (e.g., Red, NIR) and derived indices (e.g., NDVI, NDWI).

Each univariate time series is divided into overlapping patches of fixed length P, with stride $S \leq P$. For instance, a 4-year series with $L = 104$ timesteps and $P = 16, S = 8$ produces approximately 12 patches per band (Zerveas et al., 2021; Zhou et al., 2021). Each patch is treated as a "token" and serves as the unit of input to the transformer. Tokenization is done independently for each band.

2.2 Masked Modeling Objective

Inspired by masked language modeling in NLP, we randomly mask a fixed fraction (e.g., 40%) of the patch tokens in each channel (Devlin et al., 2019). The model is trained to reconstruct the masked patches using context from the surrounding (visible) patches. This patch-wise reconstruction enables the model to capture both short-range dynamics within a patch and long-range seasonal or multi-year structure across the full sequence.

2.3 Transformer Encoder

The core of the model is a 600K-parameter Transformer encoder. It processes each band's patch sequence independently using shared weights across all channels. This encourages spectral channel independence and prevents the mixing of noise distributions. The encoder outputs contextualized patch embeddings of shape (X, N, D), where X is the number of pixels, N is the number of patches per sequence, and D is the embedding dimension.

After encoding, we obtain a D-dimensional embedding for each patch. To derive a single representation for the entire time series of a pixel, we aggregate the patch embeddings using mean pooling or a special CLS token. These pooled pixel embeddings are then used for downstream tasks such as classification, clustering, or unsupervised change detection via distance metrics (e.g., L2, cosine).

3 Experiments

3.1 Dataset and Pre-Training

We evaluated our approach on a large-scale dataset derived from Sentinel-2 biweekly composites spanning five years (2020–2024). Specifically, we sampled 50,000 image "chips," each covering a 1 km × 1 km area (corresponding to 100 × 100 pixels at Sentinel-2's native 10 m resolution), randomly distributed across diverse land cover types throughout the African continent. Each chip contributes a multispectral, biweekly time series comprising approximately 130 temporal observations per pixel. To preprocess the data, we merged multiple Sentinel-2 acquisitions within each biweekly window and computed the median reflectance to mitigate cloud and noise artifacts. We retained 10 spectral bands and indices ['Green', 'Blue', 'Red', 'NIR', 'SWIR1', 'SWIR2', 'NDMI', 'NDWI', 'CI', 'NDVI'] to construct the final pixel level time series dataset.

Over the course of training, the loss steadily declined (from 0.1715 to 0.0232), while validation loss stabilized around 0.0864–0.0211, indicating limited overfitting and good reconstruction.

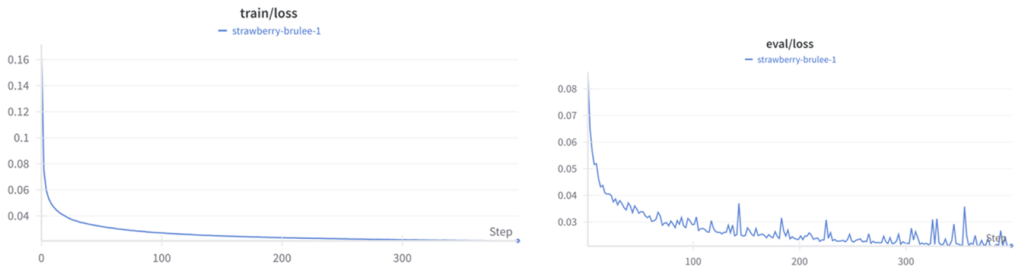

Figure 1: Training and validation loss curves over the course of self-supervised pretraining.

3.2 Evaluation and Downstream Tasks

We evaluated the learned representations on a downstream land cover classification task, as shown in Figure 2. To provide supervision, we used the ESRI Land Use and Land Cover (LULC) dataset as a source of "weak labels" across our study regions. Although these labels are not pixel-accurate and may include some noise, their broad spatial coverage and accessibility make them suitable for large-scale benchmarking. Using the pooled pixel embeddings produced by the model, we trained a lightweight classifier and observed strong diagonal dominance in the resulting confusion matrix. This indicates that the learned self-supervised representations are discriminative enough to separate major land cover types, even in the absence of precise annotations.

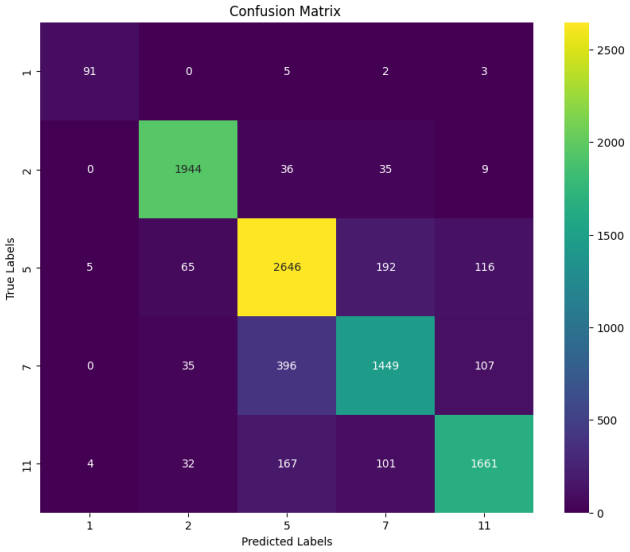

Figure 2: Confusion matrix for land-cover classification using pooled pixel embeddings and ESRI LULC labels.

To further examine the structure and quality of the representations learned during the self-supervised pretraining, we applied Principal Component Analysis (PCA) to the pooled pixel embeddings. The resulting 2D projection, shown in Figure 3, reveals coherent clusters and smooth transitions in the embedding space. These emergent patterns often correspond to the underlying spatial or temporal structure in the data. The clear separability between clusters, despite the lack of supervision, suggests that the model captures both intra-class variation and inter-class distinctions. These findings support the effectiveness of the learned embeddings for downstream geospatial and environmental monitoring tasks.

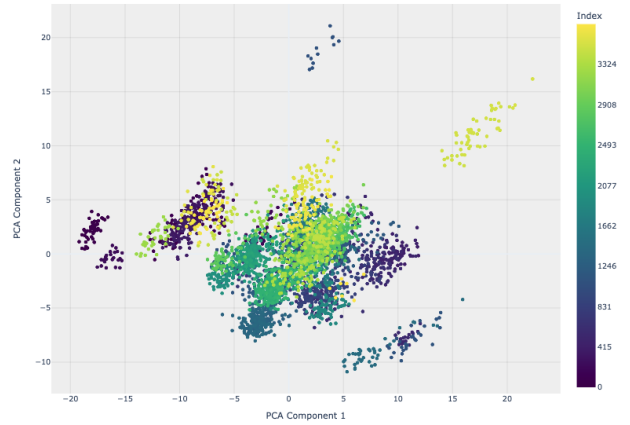

Figure 3: PCA projection of pixel embeddings, revealing structure and smooth clustering.

4 Conclusion

In this ongoing work, we present a patch-based masking approach for self-supervised learning on remote-sensing pixel time series, showing that even a relatively compact 600K-parameter Transformer effectively learns both short-range and global temporal semantics in unlabeled remote sensing data. Future work will focus on foundation-scale training, extending our scheme to larger regions and additional sensor modalities (e.g., SAR, weather, etc.).

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. URL https://arxiv.org/abs/1810.04805.

Menghao Huo, Kuan Lu, Yuxiao Li, and Qiang Zhu. Ct-patchtst: Channel-time patch time-series transformer for long-term renewable energy forecasting. *arXiv preprint arXiv:2501.08620*, 2025. doi: 10.48550/arXiv.2501.08620. URL https://arxiv.org/abs/2501.08620.

Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Parker Van-Valkenburgh, Steven A Wernke, and Yuankai Huo. Ai foundation models in remote sensing: A survey. In *arXiv preprint arXiv:2408.03464*, 2024.

Yuqi Nie, Nam Nguyen, Haoyi Phan, Abhijeet Chandra, Yuhang Li, and Chengkai Guo. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. In *arXiv preprint arXiv:2412.02732*, 2025.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. In *arXiv preprint arXiv:2304.14065*, 2024.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021. URL https://arxiv.org/abs/2010.02803.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021. doi: 10.48550/arXiv.2012.07436. URL https://arxiv.org/abs/2012.07436.