Towards Scalable Foundation Model for Multi-modal and Multi-spectral Geospatial Data

Haozhe Si [1] **Yuxuan Wan** [1] **Minh Do** [1] **Deepak Vasisht** [1] **Han Zhao** [1] **Hendrik F. Hamann** [2]

Abstract

Geospatial raster (imagery) data, such as that collected by satellite-based imaging systems at different times and spectral bands, hold immense potential for enabling a wide range of high-impact applications. Recent work has adapted existing self-supervised learning approaches for such geospatial data. However, they fall short of scalable model architectures, leading to inflexibility and computational inefficiencies when faced with an increasing number of channels and modalities. To address these limitations, we introduce our Low-rank Efficient Spatial-Spectral Vision Transformer (LESS ViT) architecture. We pretrain LESS ViT using a Multi-spectral Masked Autoencoder paradigm, and evaluate the resulting performance on our constructed GFM-Bench, a comprehensive benchmark for such geospatial raster data. Experimental results demonstrate that our proposed method achieves competitive performance against state-of-the-art multi-modal geospatial foundation models while outperforming them on cross-satellite generalization tasks with higher computational efficiency. The flexibility and extensibility of our framework make it a promising direction for future geospatial data analysis tasks that involve a wide range of modalities and channels.

1. Introduction

Geospatial data provides location-specific, timestamped information about the Earth's surface. The rapid development and proliferation of satellite-based imaging systems have led to a significant increase in geospatial raster (e.g., imagery) data collection, offering valuable insights into various aspects of our planet. Geospatial raster data is inherently multi-modal, integrating observations from diverse sensing systems such as optical and radar satellites. Each modality captures distinct data dimensions through multiple channels (e.g., spectral bands, polarizations), while introducing complexities from multi-temporal observations, non-ideal imaging conditions, and varying spatial resolutions.

Self-supervised learning (SSL) allows models to benefit from the vast amounts of unlabeled geospatial data and learning useful representations. Recent works on geospatial foundation models have attempted to adapt existing SSL paradigm to geospatial datasets using various strategies. However, although these approaches achieved empirical success, their underlying architectures and objectives remain largely the same as those designed for natural images and thus do not fundamentally suitable to fully capture the spatial, spatial-channel and inter-channel relations of geospatial data. Therefore, developing model architectures that explicitly encode these distinctive relationships in geospatial data and while efficiently scaling to thousands of spectral channels would significantly advance existing approaches that predominantly leverage spatial features.

In this work, we design a novel model architecture, Low-rank Efficient Spatial-Spectral Vision Transformer (LESS ViT), which compute the spatial-spectral attention of multi-spectral geospatial data efficiently. For pretraining, we extend the Masked Autoencoder framework to Multi-spectral MAE (Multi-MAE), which introduces a more challenging pretraining objective that encourages learning of inter-channel relationships. To standardize evaluation protocols, we construct GFM-Bench with proper validation splits and consistent metrics across diverse geospatial tasks. Extensive experiments demonstrate the effectiveness of our proposed architecture and pretraining strategy. Code and project page available at `https://uiuctml.github.io/GeospatialFM/`.

2. Low-rank Efficient Spatial-Spectral ViT

In this section, we introduce our Low-rank Efficient Spatial-Spectral (LESS) ViT architecture. Specifically, we elaborate the three key components of the framework: the multi-spectral patch embedding block, the LESS Attention Block and the Perception Field Mask.

[1]University of Illinois Urbana-Champaign, IL, USA [2]IBM Research, NY, USA. Correspondence to: Haozhe Si <haozhes3@illinois.edu>.

TerraBytes - Towards global datasets and models for Earth Observation, 42^{nd} International Conference on Machine Learning.

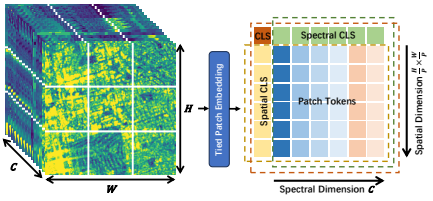

Figure 1. **Multi-spectral Patch Embedding.** Multi-spectral images, with dimensions $C \times H \times W$, are embedded into spatial-spectral tokens through the Tied Patch Embedding Layer. We then prepend the Spatial, Spectral and global `[CLS]` tokens to the resulting patch tokens.

2.1. Multi-spectral Patch Embedding

Multi-spectral images can contain tens to thousands of channels, which exhibit strong physical correlations that must be effectively leveraged. To exploit these rich spectral information in subsequent attention blocks, we adopt a Tied Patch Embedding Layer that maintains spectral fidelity by explicitly embedding each channel's information, and incorporate a continuous positional-channel embedding to capture both spatial and spectral relationships.

Tied Patch Embedding Layer. Given a multi-spectral image with dimensions $C \times H \times W$, where C denotes the number of channels, the tied patch embedding layer (Bao et al., 2023) partitions the image into $C \times \frac{H}{P} \times \frac{W}{P}$ patches of size $P \times P$. A shared learnable projection matrix $W \in \mathbb{R}^{P^2 \times D}$ transforms patches from each channel into D-dimensional tokens. This weight-sharing mechanism across channels (Ghiasi et al., 2022) ensures channel-independence, making the architecture adaptable to geospatial data with varying spectral dimensions. The resulting spatial-spectral tokens have dimension $\mathbb{R}^{N \times C \times D}$, where $N = \frac{H}{P} \times \frac{W}{P}$.

Continuous Positional-Channel Embeddings. To ensure positional consistency across datasets with varying spatial resolutions, we compute absolute geographic distances:

$$\begin{aligned} \mathrm{PE}_{(x,r,p,2i)} &= \sin(xrp/10000^{2i/d}), \\ \mathrm{PE}_{(x,r,p,2i+1)} &= \cos(xrp/10000^{2i/d}), \end{aligned} \tag{1}$$

where x denotes the grid index, p is the patch size, r represents the image spatial resolution in meters per pixel, d is the model's embedding dimension, and $i \in \{0, 1, \ldots, \lfloor d/2 \rfloor - 1\}$. For the spectral axis, we encode the central wavelength λ of each multi-spectral band using a similar formulation:

$$\begin{aligned} \mathrm{PE}(\lambda, 2i) &= \sin(\lambda/10000^{2i/d}), \\ \mathrm{PE}(\lambda, 2i+1) &= \cos(\lambda/10000^{2i/d}), \end{aligned} \tag{2}$$

This physics-informed embedding enables the model to arbitrary spectral bands, as it maps channels to a continuous spectral space rather than treating them as discrete indices. Finally, we sum the spatial and spectral embeddings to form our continuous positional-channel embedding that jointly encodes both geographic distances and spectral wavelength.

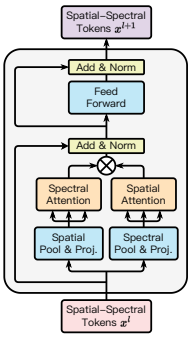

Figure 2. **LESS Attention Block.** An illustration of the LESS Attention Block, which decomposes spatial and spectral attention computations and approximates the full spatial-spectral attention through a Kronecker product of the individual attention maps.

2.2. Low-rank Efficient Spatial-Spectral Attention

Given the spatial-spectral tokens $X \in \mathbb{R}^{N \times C \times D}$, we would like to apply the attention mechanism to capture the correlations of the spatial and spectral patches. A straightforward yet inefficient approach (Bao et al., 2023) is to flatten the spatial-spectral tokens into $\bar{X} \in \mathbb{R}^{NC \times D}$ and apply the standard attention mechanism. The computational complexity of this approach scales quadratically with the number of channels and tokens ($O(N^2C^2)$), making it infeasible for geospatial data with a large number of channels or tokens. To address this limitation, we propose the Low-rank Efficient Spatial-Spectral Vision Transformer (LESS ViT). LESS ViT consists of multiple LESS attention blocks specifically designed for spatial-spectral tokens. An illustration of the blocks is shown in Figure 2. The computation complexity of LESS ViT is reduced to $O(NC)$, which scales linearly with the number of spatial-spectral tokens.

To efficiently model spatial-spectral interactions, our LESS attention block approximates the full spatial-spectral attention matrix using a Kronecker product of separate spatial and spectral attention matrices. Specifically, the block first decomposes input tokens X into spatial tokens X_S and spectral tokens X_C. Then, the spatial attention matrix A_S and the spectral attention matrix A_C are calculated separately using X_S and X_C, along with their respective value matrices V_S and V_C. Since A_S and A_C represent convex combinations of spatial and spectral dimensions respectively, their Kronecker product $A = A_C \otimes A_S$ yields a convex combination over the joint spatial-spectral dimensions. Leveraging the mixed-product property, we efficiently obtain a low-rank approximation of the full attention computation:

$$Y := \sum_{i=1}^{r} (A_C^i \otimes A_S^i)(V_C \otimes V_S) = \sum_{i=1}^{r} Y_C^i \otimes Y_S^i, \tag{3}$$

where $Y_S^i = A_S^i V_S \in \mathbb{R}^{N \times d_1}$ and $Y_C^i = A_C^i V_C \in \mathbb{R}^{C \times d_2}$, $\forall i \in \{1, \cdots r\}$. This approach avoids explicitly construct-

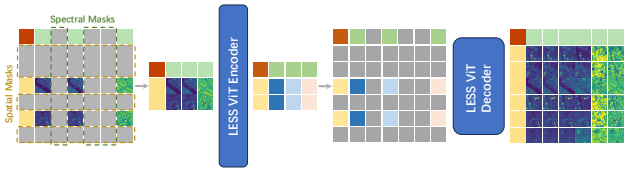

Figure 3. **Multi-MAE.** Multi-MAE employs a LESS ViT encoder-decoder architecture. The framework incorporates decoupled spatial and spectral masking to create a more challenging self-supervised pretraining objective.

ing the full attention matrix A, thereby reducing computational complexity. Note that r in Equation (3) is a hyperparameter for rank control. By adjusting the rank, we can increase the capacity of the attention map while maintaining the same computational complexity order.

The LESS attention block offers significant efficiency advantages compared to the previous spatial-spectral attention approach (Bao et al., 2023), which directly applies attention to the reshaped tokens $\bar{X} \in \mathbb{R}^{NC \times D}$. It reduces the computational complexity from $O(N^2C^2D)$ to $O(rN^2d_1 + rC^2d_2 + rNCD)$, where $d_1d_2 = D$ and $r \ll \min(N, D)$ is a small constant.

2.3. Perception Field Mask

To explicitly model spatial autocorrelation in geospatial data, we introduce the Perception Field Mask. The Perception Field Mask constrains the spatial attention computation by allowing each token to attend only to patches within a specified distance threshold. Consequently, This distance-based masking mechanism offers two key advantages: (1) it provides a tunable hyperparameter to control the locality of attention computation, aligning with Tobler's law, and (2) it enables the model to process images of varying sizes without downsampling, as the attention field remains spatially consistent regardless of resolution.

3. Multi-spectral Masked Autoencoder

We extend the Masked Autoencoder (MAE) framework and propose the Multi-spectral Masked Autoencoder (Multi-MAE), which decouples spatial and spectral masking. An illustration is shown in Figure 3. During training, we randomly mask 75% of spatial patches and 50% of spectral channels before encoding. The decoder then reconstructs the complete multi-spectral image at pixel level from these partially observed tokens. Importantly, we apply identical spatial masks across all unmasked channels, similar to the tube masking strategy in (Tong et al., 2022). This approach encourages the model to learn intrinsic spatial and spectral correlations in geospatial data, as it cannot rely on positional information from other channels to reconstruct masked spatial patches.

4. GFM-Bench

Several datasets do not include validation sets, forcing prior work to tune hyperparameters on test data. To address this limitation, we introduce GFM-Bench, implemented using the HuggingFace framework and providing standardized evaluation protocols. The current version of GFM-Bench consists three classification tasks (EuroSAT (Helber et al., 2019), BigEarthNet (Sumbul et al., 2019) and So2Sat (Zhu et al., 2020) and four segmentation tasks (SegMunich (Hong et al., 2024), DFC2020 (Yokoya et al., 2020), MARIDA (Kikaki et al., 2022), NLCD-L (Stewart et al., 2024)). All datasets are derived from data of the Sentinel constellation except NLCD-L, which uses Landsat data. For datasets without validation splits, we either allocate 10% of the training data for validation or utilize alternate versions that include validation sets. We implement consistent evaluation metrics to ensure fair hyperparameter selection across models. GFM-Bench enforces hyperparameter tuning on validation sets and performance reporting on test sets.

5. Experiments

We pretrain a LESS ViT-Base model using Multi-MAE on the Senetinel 1&2 data from SSL4EO-S12 dataset (Wang et al., 2023). We evaluate our models and the publicized checkpoints from the baseline methods on GFM-Bench and present quantitative experimental results to demonstrate out competitive performance against state-of-the-art approaches.

5.1. Multispectral Optical Experiments

Our experimental results for classification and segmentation tasks are summarized in Table 1. LESS ViT demonstrates competitive performance across most benchmarks compared to existing approaches. Among ViT-Base-sized models, our approach achieves the second-highest average performance, indicating robust generalization across diverse downstream tasks. Notably, LESS ViT outperforms several ViT-Large-sized baselines on specific benchmarks, despite its more compact architecture. The linear probing (LP) results demonstrate that LESS ViT learns transferable representations for multi-spectral geospatial data, while the fine-tuning (FT) results highlight its strong task-specific adaptation capabilities. However, we observe performance gaps on So2Sat(Zhu et al., 2020) and DFC2020 (Yokoya et al., 2020) datasets. These gaps arise from distribution shifts between training and test sets, motivating future work to enhance training robustness through improved self-supervision objectives.

Table 1. **Quantitative results on seven benchmarks under Fine-tuning (FT) and Linear Probing (LP).** We report Top 1 accuracy for classification tasks, mean Average Precision (mAP) for multi-label classification tasks, and mean Intersection over Union (mIoU) for segmentation tasks. * indicates only 10% of the training and validation sets are used, following previous works. We also report the average performance of each method on all benchmarks. **Bold** and underlined values indicating the highest and second-highest results. The bottom two rows show ViT-Large models, which serve as references and are not directly compared with the ViT-Base approaches.

		EuroSAT Top 1 Acc.		BigEarthNet* mAP		So2Sat* Top 1 Acc.		SegMunich mIoU	DFC2020 mIoU	MARIDA mIoU	*Avg.*
Method	Backbone	FT	LP	FT	LP	FT	LP	FT	FT	FT	
SatMAE (Cong et al., 2022)	ViT-B	98.78	**96.04**	85.84	78.69	64.97	64.65	**44.87**	**52.84**	54.33	**71.22**
CROMA (Fuller et al., 2024)	ViT-B	**98.83**	95.87	**87.57**	**84.90**	**66.53**	**65.04**	39.61	49.48	43.04	70.10
SpectralGPT (Hong et al., 2024)	ViT-B	97.94	90.57	83.78	73.29	61.63	57.49	44.73	48.23	44.72	66.93
Ours	LESS ViT-B	98.06	95.12	86.08	82.94	63.25	64.66	42.29	45.60	**55.64**	70.40
Scale-MAE (Reed et al., 2023)	ViT-L	98.78	96.41	84.66	73.69	66.48	60.84	44.84	48.75	41.12	68.51
SatMAE++ (Noman et al., 2024)	ViT-L	98.91	94.61	85.89	79.08	65.18	60.40	45.86	52.02	59.82	71.31

Table 2. **Cross-Satellite Generalization to Landsat and Model Efficiency.** We evaluate architectures for cross-satellite generalization and computational efficiency on NLCD-L (Stewart et al., 2024), a 20-channel Landsat segmentation dataset. To enable direct comparison, we normalize both FLOPs and wall-clock times relative to LESS ViT's baseline measurements.

Dataset	**Architecture**	**Backbone**	**#Param.**	**Fine-Tuning Time**	**# FLOPs**	**Inference Time**	**mAP**
NLCD-L (Stewart et al., 2024)	SatMAE (Cong et al., 2022)	ViT-B	86.1M	×0.3	×0.6	×0.3	18.05
	Channel-ViT (Bao et al., 2023)	Channel-ViT-B	85.4M	×2.6	×3.1	×3.6	10.35
	Ours	LESS ViT-B	83.2M	×1.0	×1.0	×1.0	**24.31**

5.2. Cross-Satellite Generalization

We consider cross-satellite generalization as a critical ability of future geospatial **foundation** models. To demonstrate the flexibility of our LESS ViT architecture in handling satellites with varying channel counts without architectural modifications, we evaluate our model on the NCLD-L dataset from GFM-Bench. We construct this dataset by combining optical data from Landsat 7 and Landsat 8-9 from SSL4EO-L (Stewart et al., 2024), resulting in a 20-channel geospatial dataset that exceeds Sentinel-2's channel count. We compare LESS ViT against two baseline architectures: SatMAE (Cong et al., 2022), a ViT-based model, and Channel-ViT (Bao et al., 2023), which explicitly models spatial-spectral attention. We fine-tune the three Sentinel-pretrained base models on NCLD-L for 10 epochs, with results shown in Table 2. Beyond channel count differences, Landsat features lower spatial resolution (30.0 meters/pixel) compared to Sentinel (10.0 meters/pixel). From the results we can see that our model architecture generalizes to these variations better than the previous architectures.

5.3. Model Efficiency

To evaluate LESS ViT's efficiency, we measure fine-tuning and inference wall-clock time, parameter counts, and floating point operations (FLOPs) compared to ViT-based SatMAE (Cong et al., 2022) and Channel-ViT (Bao et al., 2023) on NLCD-L (Dewitz et al., 2021). As shown in Table 2, both LESS ViT and Channel-ViT reduce parameter counts compared to ViT through their tied patch embedding layers, which share embedding weights across spectral channels. LESS ViT achieves the lowest parameter count through its low-rank attention module. ViT demonstrates the fastest fine-tuning and inference times by collapsing the spectral dimension during patch embedding, though it omits the spectral attention. In contrast, Channel-ViT's explicit spatial-spectral attention computation leads to the highest FLOPs and lowest computational efficiency. Despite computing spatial-spectral attention explicitly, Channel-ViT fails to benefit from this approach as it does not outperform ViT-based models, constrained by the inevitable random channel masking during training. Conversely, our LESS ViT approximates spatial-spectral attention more efficiently, eliminating the need for channel masking and enabling better training data utilization.

6. Limitations and Future Works

While LESS ViT demonstrates competitive performance with improved computational efficiency, extending the approach to extra dimensions (e.g., temporal dimension) remains a challenge. Although LESS attention can accommodate additional dimensions, this expansion results in reduced embedding dimensions d_n per dimension. We propose exploring model scaling strategies through enhanced embedding dimension allocation in future research. Additionally, our current approach is limited to raster representations, while the remote sensing community possesses rich domain knowledge typically represented in vector formats, such as digital elevation models and slope models. Integrating these vector-based domain knowledge with raster (imagery) data remains an unresolved research challenge for future methodological advancements. Nevertheless, this work advances architectural design for hyperspectral data processing and deepens our understanding of multidimensional correlations. The proposed framework establishes a foundation for future developments in Earth Observation tasks and geospatial data analysis in the remote sensing community.

References

Bao, Y., Sivanandan, S., and Karaletsos, T. Channel vision transformers: An image is worth 1 x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Dewitz, J. et al. National land cover database (nlcd) 2019 products. *US Geological Survey*, 10:P9KZCM54, 2021.

Fuller, A., Millard, K., and Green, J. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., Wilson, A. G., and Goldstein, T. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitsos, D. E., and Karantzalos, K. Marida: A benchmark for marine debris detection from sentinel-2 remote sensing data. *PloS one*, 17(1):e0262247, 2022.

Noman, M., Naseer, M., Cholakkal, H., Anwar, R. M., Khan, S., and Khan, F. S. Rethinking transformers pre-training for multi-spectral satellite imagery. In *CVPR*, 2024.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Stewart, A., Lehmann, N., Corley, I., Wang, Y., Chang, Y.-C., Ait Ali Braham, N. A., Sehgal, S., Robinson, C., and Banerjee, A. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36, 2024.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE, 2019.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

Yokoya, N., Ghamisi, P., Hänsch, R., and Schmitt, M. 2020 ieee grss data fusion contest: Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157, 2020.

Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., et al. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3): 76–89, 2020.