TerraMAE: Learning Spatial-Spectral Representations from Hyperspectral Earth Observation Data via Adaptive Masked Autoencoders

Tanjim Bin Faruk
Colorado State University
Fort Collins, Colorado, USA
tanjim@colostate.edu

Abdul Matin
Colorado State University
Fort Collins, Colorado, USA
amatin@colostate.edu

Shrideep Pallickara
Colorado State University
Fort Collins, Colorado, USA
Shrideep.Pallickara@colostate.edu

Sangmi Lee Pallickara
Colorado State University
Fort Collins, Colorado, USA
Sangmi.Pallickara@colostate.edu

Abstract

Masked Autoencoders struggle with hyperspectral satellite imagery containing 200+ spectral bands, as uniform masking across all channels obscures critical spatial-spectral relationships. We introduce TerraMAE, which employs an adaptive channel grouping strategy to organize bands into statistically coherent groups with independent masking. Together with a customized loss function, this data-driven grouping strategy enables TerraMAE to learn robust spatial-spectral representations from unlabeled HSI. Experiments demonstrate that TerraMAE significantly outperforms baseline Masked Autoencoder and supervised ResNet-50 on soil texture prediction, achieving 15.7% and 6.6% lower error, respectively.

CCS Concepts

• **Computing methodologies** → **Learning latent representations**; **Hyperspectral imaging**; *Neural networks*; • **Information systems** → *Geographic information systems.*

Keywords

Hyperspectral Satellite, Geo AI, Masked Autoencoders, Deep Learning, Self-supervised Learning, Remote Sensing

ACM Reference Format:
Tanjim Bin Faruk, Abdul Matin, Shrideep Pallickara, and Sangmi Lee Pallickara. 2025. TerraMAE: Learning Spatial-Spectral Representations from Hyperspectral Earth Observation Data via Adaptive Masked Autoencoders. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3748636.3762770

1 Introduction

Hyperspectral satellite imagery (HSI) captures Earth's surface in 200+ contiguous spectral bands, enabling more detailed analysis of soil properties, vegetation health, and land cover compared to conventional 3–11 band multispectral imagery. However, extracting meaningful representations from HSI remains challenging due to the scarcity of labeled data and the complex spatial-spectral dependencies across hundreds of channels.

Masked Autoencoders (MAEs) offer a promising self-supervised approach for learning from unlabeled satellite data, with recent adaptations like SatMAE [3] showing success on multispectral imagery. However, these methods face scalability challenges with HSI's high dimensionality. When MAEs apply uniform masking across 200+ channels, they can obscure important spatial-spectral correlations—geographic regions may become entirely masked across all bands, reducing the contextual information available for reconstruction. While SatMAE addressed this for 13-band Sentinel-2 data using fixed wavelength-based channel groups, such static partitions may not capture the complex spectral relationships in HSI, where meaningful correlations exist between non-adjacent bands. Recent variants like S2MAE [7] also explore spatial-spectral masking, but focus on classification-specific pretraining rather than general-purpose representations. SpectralGPT [5] and other 3D MAE variants target multispectral rather than hyperspectral data.

Recent HSI foundation models have pursued scale over architectural innovation. SpectralEarth [1] trained on 538k patches but lacks publicly available model weights, while HyperSIGMA [9] and DOFA [10] use proprietary datasets and face reproducibility challenges. The lack of public weights and standardized benchmarks prevents direct comparison, motivating our focus on architectural innovations that work with smaller, accessible datasets. Our methodology is broadly in the area of science-informed learning for spatiotemporally evolving phenomena [4, 6, 8].

1.1 Our Approach

We address the question: How can self-supervised MAEs be adapted to effectively pretrain models on HSI given its complex spatial-spectral structure? We introduce TerraMAE with two key innovations: (1) adaptive channel grouping using the Spectral Comparison Index (SCI), which clusters channels based on spatial-spectral similarity rather than fixed wavelengths, applying independent masks within each group to preserve spatial context; and (2) a composite reconstruction loss combining Mean Absolute Error with SSIM and SID, jointly optimizing for spatial structure and spectral fidelity. Our

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3762770

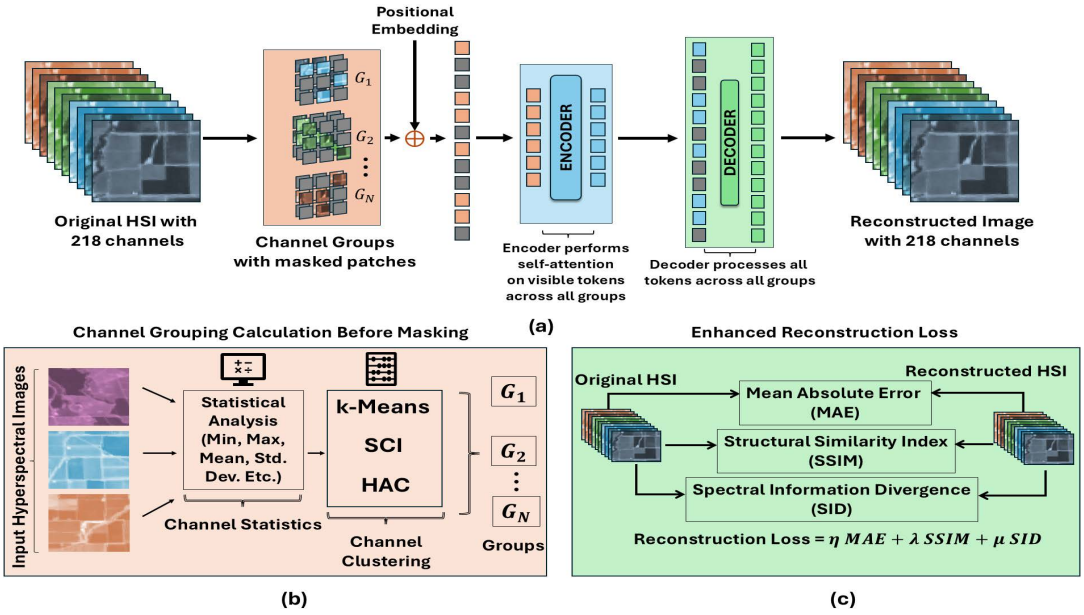

Figure 1: TerraMAE Architecture.

experiments demonstrate that TerraMAE consistently outperforms baseline MAE and supervised ResNet-50 across reconstruction and downstream tasks.

2 Methodology

The overall structure of TerraMAE is presented in Figure 1.

2.1 Adaptive Channel Grouping with Independent Masking

Standard MAEs apply uniform masks across all channels, which works for RGB images but becomes problematic for HSI with 200+ bands. Uniform masking can result in entire spatial regions being masked across all channels, eliminating critical contextual information needed for reconstruction–especially challenging where subtle spectral signatures encode essential geospatial information.

TerraMAE addresses this challenge by grouping spectrally similar channels and applying masks independently within each group while maintaining the predefined masking ratio. This preserves spatial context across the spectral dimension while reducing computational complexity.

We introduce the Spectral Comparison Index (SCI) to quantify channel similarity:

$$SCI_{i,j} = 1 - \frac{|I_i - I_j|}{I_i + I_j + \epsilon} \tag{1}$$

where I_i and I_j are the reflectance values of bands i and j respectively, and ϵ prevents division by zero. The SCI values range from 0 to 1, with higher values indicating greater similarity between the bands.

For each channel pair, we compute SCI at each spatial location, then aggregate into a single similarity score:

$$SCI_{prod} = SCI_{\mu} \times (1 - SCI_{\sigma}) \tag{2}$$

where SCI_{μ} is the mean and SCI_{σ} is the standard deviation across the spatial map. The stability term $(1 - SCI_{\sigma})$ down-weights channel pairs with high spatial variability, favoring groups with consistent spatial-spectral patterns. Channels are then clustered into a predefined number of groups (5 in our experiments, balancing reconstruction quality with computational efficiency as more groups improve performance but increase training time significantly) based on SCI_{prod} scores. This reflectance-based grouping captures physical surface properties that remain consistent across spatial regions, making it more robust than intensity-based metrics for geospatial applications.

We compare SCI against: (1) k-Means and HAC clustering using statistical features (mean, std, range, etc.), (2) VNIR-SWIR grouping (2 fixed wavelength groups), and (3) Soil-Reflectance grouping (5 domain-based groups) [2].

2.2 Enhanced Loss Function

Pixel-level losses like Mean Absolute Error fail to capture spatial structure and spectral relationships critical for HSI. We augment Mean Absolute Error with two components:

- **Spatial Coherence via SSIM**: To preserve field boundaries and spatial patterns, we incorporate the Structural Similarity Index (SSIM):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

where μ_x, μ_y are local means, σ_x, σ_y are standard deviations, σ_{xy} is covariance, and c_1, c_2 are stability constants.

Table 1: Reconstruction results on California test sets (5 spectral groups). Me. Abs. Err. refers to Mean Absolute Error

Setup	Grouping	Loss Function	Test Set 1			Test Set 2 (Unseen)		
			Me. Abs. Err. ↓	PSNR ↑	SSIM ↑	Me. Abs. Err. ↓	PSNR ↑	SSIM ↑
Baseline	×	Me. Abs. Err.	0.0172	27.12	0.4227	0.0181	27.35	0.4092
TerraMAE	**SCI**	**Me. Abs. Err. + SSIM + SID**	**0.0047**	**37.47**	**0.9112**	**0.0051**	**37.55**	**0.9083**

SSIM is normalized to $[0, 1]$ as $\text{SSIM}_N = \frac{(1-\text{SSIM})}{2}$ for loss computation.
While SSIM was designed for visible imagery, it effectively captures spatial structure regardless of wavelength, as validated by improved downstream performance in Section 4.2.

- **Spectral Fidelity via SID**: To maintain spectral signatures crucial for material identification, we use Spectral Information Divergence (SID):

$$\text{SID}(x, y) = \sum_{i=1}^{C} \left(p_i \log \left(\frac{p_i}{q_i} \right) + q_i \log \left(\frac{q_i}{p_i} \right) \right) \tag{4}$$

where p_i and q_i are normalized spectral components. To scale comparably with other loss terms, SID is scaled as $\text{SID}_N = 1 - e^{-\alpha \times \text{SID}}$.

Our final loss combines all three components:

$$\mathcal{L} = \eta \times \text{MeanAbsoluteError} + \lambda \times \text{SSIM}_N + \mu \times \text{SID}_N \tag{5}$$

where the weights $(\eta, \lambda, \mu) = (0.7, 0.15, 0.15)$ were selected via hyperparameter sweep to balance pixel-wise accuracy with spatial and spectral fidelity.

3 Pretraining Experiments and Results

We collected 22,078 EnMAP hyperspectral image tiles (218 bands after removing 6 water vapor absorption bands) from California's agricultural regions, excluding areas with >10% cloud cover, snow, or mountainous terrain. Each tile covers 2km × 2km at 30m resolution, acquired between 2022–2024. We split the data into 12,466 training, 2,078 validation, and 7,534 test tiles. The test set was further divided: Test Set 1 (TS1) contains 6,234 tiles, while Test Set 2 (TS2) comprises 1,300 spatially disjoint tiles with no geographic overlap with training/validation data to assess intra-state generalization.

We pretrain TerraMAE using a ViT-Large backbone for 300 epochs with a 75% masking ratio and 5 spectral groups. The baseline uses an identical architecture without spectral grouping or enhanced loss. During downstream tasks, pretrained encoders are frozen while ResNet-50 trains from scratch.

3.1 Evaluation Metrics

We evaluate reconstruction using Mean Absolute Error, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM).

3.2 Reconstruction Performance

Table 1 shows TerraMAE substantially outperforms the baseline on reconstruction metrics. The 73% reduction in MAE and 10dB PSNR improvement demonstrate effective spatial-spectral learning. Performance remains strong on geographically disjoint TS2, validating robust generalization.

4 Transfer Learning: Soil Texture Prediction

To evaluate TerraMAE's learned representations, we focus on soil texture prediction–estimating sand, silt, and clay proportions critical for precision agriculture. We freeze the pretrained encoders and train only a lightweight CNN decoder (220k parameters) for downstream adaptation, comparing against baseline MAE and ResNet-50 trained from scratch.

4.1 Downstream Dataset

We use POLARIS 30m resolution soil data for California as ground truth, predicting sand and silt fractions (clay is derived as 100% minus sand and silt percentages).

4.2 Results

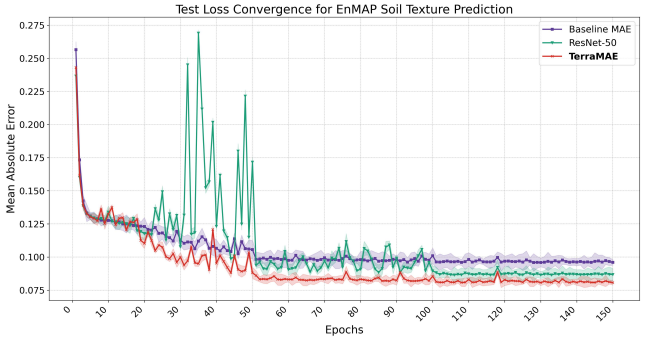

Figure 2: Test loss convergence over 150 epochs for soil texture prediction using different encoder initializations.

Table 2: EnMAP-POLARIS Soil Texture Prediction Results (Mean Absolute Error ± Standard Deviation over 5 runs).

Method	Mean Absolute Error ± Std. Dev. ↓
ResNet-50	0.0863 ± 0.0032
Baseline MAE + CNN	0.0956 ± 0.0041
TerraMAE + CNN	**0.0806 ± 0.0023**

Table 2 shows TerraMAE achieves 6.6% lower error than supervised ResNet-50 and 15.7% improvement over baseline MAE,

demonstrating that HSI-specific pretraining enables a compact decoder to outperform deeper networks.

TerraMAE not only reaches the lowest Mean Absolute Error but also converges more smoothly (Figure 2); important for geospatial regression tasks where training instability impacts map prediction quality.

Table 3 ablates grouping strategies and loss functions, confirming SCI consistently outperforms alternatives (k-Means, HAC, domain-based SR). The combination of SCI grouping with enhanced loss (Me. Abs. Err. + SSIM + SID) achieves the best performance; adaptive spectral grouping and composite loss functions enable TerraMAE to learn transferable representations from unlabeled HSI.

Table 3: Effect of grouping strategies on soil texture prediction. Me. Abs. Err. denotes Mean Absolute Error.

Grouping Strategy	Loss Function	Me. Abs. Err. ↓
×	Me. Abs. Err.	0.1027
×	Me. Abs. Err. + SSIM + SID	0.0956
SR	Me. Abs. Err.	0.0938
SR	Me. Abs. Err. + SSIM + SID	0.0907
k-Means	Me. Abs. Err.	0.0879
k-Means	Me. Abs. Err. + SSIM + SID	0.0843
HAC	Me. Abs. Err.	0.0855
HAC	Me. Abs. Err. + SSIM + SID	0.0822
SCI	**Me. Abs. Err.**	**0.0828**
SCI	**Me. Abs. Err. + SSIM + SID**	**0.0806**

5 Conclusion

TerraMAE is a self-supervised framework that adapts MAEs to HSI through two key innovations: adaptive channel grouping based on spatial-spectral coherence rather than fixed wavelengths, and a composite loss function that jointly preserves spatial structure and spectral fidelity. TerraMAE consistently outperformed baseline MAE and supervised ResNet-50 across reconstruction and downstream tasks, demonstrating that targeted architectural design enables effective HSI representation learning even with limited data.

6 Appendix: Ablation Studies

Masking Ratio Analysis Table 4 shows that a 75% pretrain masking ratio yields the best reconstruction for TerraMAE when evaluated at the same level.

Table 4: Varying mask ratios on California Test Set 2.

Pretrain Mask Ratio (PMR)	Inference at PMR		Inference at 75% MR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
60%	38.84	0.9391	36.46	0.8961
70%	38.12	0.9245	37.10	0.9015
75%	37.55	0.9083	**37.55**	**0.9083**
80%	36.52	0.8842	36.91	0.8937
85%	34.21	0.8124	35.82	0.8684

Impact of Spectral Grouping Increasing spectral groups improves reconstruction but significantly raises training time (Table 5), reflecting the spectral granularity and computational cost trade-off.

Table 5: California Test Set 2 performance with varying channel groups, evaluated at 75% masking.

Pretrain Mask Ratio	Groups	PSNR ↑	SSIM ↑	Time/Epoch
75%	5	37.55	0.9083	∼150s
75%	10	40.13	0.9267	∼280s
75%	20	**42.47**	**0.9386**	∼670s
80%	5	36.91	0.8937	∼135s
80%	10	38.36	0.9081	∼250s
80%	20	40.02	0.9206	∼590s
85%	5	35.82	0.8684	∼127s
85%	10	37.19	0.8816	∼230s
85%	20	38.22	0.8959	∼550s

Acknowledgments

This research was supported by the National Science Foundation (1931363, 2312319), the National Institute of Food and Agriculture (2024-67021-43840), an NSF/NIFA Artificial Intelligence Institutes AI-LEAF (2023-03616) and a Clare Booth Luce Professorship.

References

[1] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. 2024. SpectralEarth: Training Hyperspectral Foundation Models at Scale. arXiv:2408.08447 [cs.CV] https://arxiv.org/abs/2408.08447

[2] S Chabrillat, Luis Guanter, Hermann Kaufmann, S Förster, Alison Beamish, Arlena Brosinsky, Hendrik Wulf, Saeid Asadzadeh, M Bochow, Niklas Bohn, et al. 2022. EnMAP science plan. *EnMAP* (2022).

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 15, 15 pages.

[4] Tanjim Bin Faruk, Abdul Matin, Shrideep Pallickara, and Sangmi Lee Pallickara. 2025. Accounting for Spatial Variability with the Histogram of Oriented Gradients Based Masking Improves Performance of Masked Autoencoder over Hyperspectral Satellite Imagery (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 28 (Apr. 2025), 29365–29367. doi:10.1609/aaai.v39i28.35253

[5] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. 2024. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5227–5244. doi:10.1109/TPAMI.2024.3362475

[6] Paahuni Khandelwal, Sangmi Lee Pallickara, and Shrideep Pallickara. 2024. DeepSoil: A Science-guided Framework for Generating High Precision Soil Moisture Maps by Reconciling Measurement Profiles Across In-situ and Remote Sensing Data. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems* (Atlanta, GA, USA) *(SIGSPATIAL '24)*. Association for Computing Machinery, New York, NY, USA, 233–246. doi:10.1145/3678717.3691261

[7] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. 2024. S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27696–27705. doi:10.1109/CVPR52733.2024.02616

[8] Abdul Matin, Paahuni Khandelwal, Shrideep Pallickara, and Sangmi Lee Pallickara. 2023. DISCERN: Leveraging Knowledge Distillation to Generate High Resolution Soil Moisture Estimation from Coarse Satellite Data . In *2023 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Los Alamitos, CA, USA, 1222–1229. doi:10.1109/BigData59044.2023.10386179

[9] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. 2025. HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025), 1–18. doi:10.1109/TPAMI.2025.3557581

[10] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. 2024. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. arXiv:2403.15356 [cs.CV] https://arxiv.org/abs/2403.15356