INFORMATION MANAGEMENT SCHOOL

Master's Degree in Geographic Information Systems and Science

Multimodal Transformer for Crop Monitoring - Literature Review

Author: Antonio Henrique Xavier da Silva (20240915)

Supervisor: Professor Roberto Henriques

1 Introduction

Modelling vegetation dynamics through remote sensing has become increasingly critical for diverse applications, ranging from precision agriculture to climate change assessment. The advent of high-resolution satellite missions, particularlly the Sentinel-2 and Landsat-8/9 constellations providing imagery up to 10m spatial resolution with 5-day revisit times, has enabled unprecedented temporal monitoring of Earth's vegetation cover. However, monitoring the complex spatio-temporal patterns inherent in satellite image time series (SITS) requires sophisticated deep learning architectures capable of capturing both local spectral features and long-range temporal dependencies.

This literature review examines the application and evolution of Vision Transformer (ViT) architectures for SITS analysis, with particular focus on their potential for vegetation index prediction. By synthetizing insights from recent advancements and critically analyzing the current state of knowledge, this review establishes a theoretical and methodological foundation for developing transformer-based architectures capable of predicting vegetation indices such as NDVI (Normalized Difference Vegetation Index) from temporal sequences of satellite imagery.

2 A Paradigm Shift in Computer Vision

2.1 The CNN Era and Its Limitations

From 2012 to 2020 Convolutional Neural Networks (CNNs) dominated computer vision, architectures such as AlexNet, VGGNet, ResNet, and EfficientNet achieved state-of-the-art performance on image recognition benchmarks (Dosovitskiy et al., 2020). Featuring strong inductive biases on its designs, including locality and translation invariance through convolutional operations, CNNs are particularly effective at capturing local features and hierarchical representations (Mehdipour et al., 2025).

However, architectural limitations constrains CNNs' effectiveness for certain vision tasks. The reliance on local receptive fields make it difficult to capture long-range dependencies without very deep networks, resulting in high computational costs and information redundancy (Khan et al., 2022). For remote sensing applications in particular, these limitations become critical when trying to model contextual relationships across large spatial extents or temporal dependencies across image sequences (Aleissaee et al., 2023).

2.2 The Transformer Revolution

The Transformer architecture, introduced by Vaswani et al., 2017, revolutionized natural language processing through self-attention mechanisms capable of capturing long-range dependencies in sequential data. Building on their work, Dosovitskiy et al., 2020 demonstrated that pure transformer architectures could match or exceed CNN performance on image classification tasks. This seminal work, titled "An Image is Worth 16×16 Words", established the foundational Vision Transformer (ViT) architecture that treats images as sequences of fixed-size patches, linearly embeds each patch, adds positional encodings, and processes the resulting sequence through standard Transformer encoder blocks.

The key innovation of ViTs lies in their ability to learn spatial relationships without explicit convolutional inductive biases, relying instead on multi-head self-attention mechanisms to dynamically focus on relevant image regions (Mehdipour et al., 2025). Critically, Dosovitskiy et al., 2020 revealed that ViTs exhibit different scaling behavior compared to CNNs, performing substantially better when pre-trained on large datasets before fine-tuning on specific tasks.

2.3 Evolution of Vision Transformer Architectures

Following the foundational ViT work, the field has rapidly diversified with specialized architectures addressing various limitations and use cases. Khan et al., 2022 provide a comprehensive taxonomy of over 60 transformer-based vision methods, categorizing them into uniform-scale designs, multi-scale hierarchical architectures, and hybrid CNN-transformer approaches. Key developments include:

- **Data-efficient training** employing distillation and advanced data augmentation to reduce the large dataset requirements of pure ViTs
- **Hierarchical designs** such as Swin Transformer and Pyramid Vision Transformer (PVT) (Liu et al., 2021; Wang et al., 2021) that process features at multiple scales, addressing the limitation of uniform-scale ViTs for dense prediction tasks
- **Efficient attention mechanisms** including local attention, windowed attention, and neighbourhood attention that reduce computational complexity from $O(N^2)$ to more manageable levels
- **Hybrid architectures** that combine CNN feature extraction with transformer-based global context modeling, balancing local inductive biases with global attention capabilities

These architectural innovations have established transformers as a competitive or superior alternative to CNNs across most computer vision tasks, with particular advantages for problems requiring long-range dependency modeling.

3 Vision Transformers in Remote Sensing

3.1 Early Adoption and Domain-Specific Challenges

The remote sensing community rapidly recognized the potential of Vision Transformers, with Bazi et al., 2021 among the first to systematically evaluate ViTs for satellite image classification. Working with three standard remote sensing datasets (UC Merced, AID, NWPU-RESISC45), they demonstrated that ViTs could achieve competitive or superior performance compared to CNNs when combined with appropriate transfer learning strategies and data augmentation techniques including CutMix, Cutout, and Mixup.

However, Aleissaee et al., 2023 identify several challenges specific to applying transformers in remote sensing contexts. Remote sensing imagery exhibits characteristics that distinguish it from natural images,

including: multi-spectral and hyperspectral characteristics extending beyond the RGB color space of typical vision datasets; varying spatial resolutions from meter-scale to kilometer-scale imagery products; atmospheric distortions affecting >60% of Earth's surface at any given time; and domain shift between pre-training datasets (typically ImageNet) and remote sensing applications.

These unique characteristics require domain-specific adaptations of transformer architectures rather than direct application of computer vision methods.

3.2 The Challenge of Temporal Modeling

Satellite image time series (SITS) present unique modeling challenges that extend beyond single-image analysis. Vegetation monitoring applications require models capable of capturing temporal dynamics including phenological cycles, growth patterns, and responses to environmental stressors. Traditional approaches using CNNs combined with RNNs or LSTMs face limitations including sequential processing bottlenecks, difficulty propagating information across long sequences, and inability to capture global temporal receptive fields.

Transformers offer compelling advantages for SITS processing through their ability to model long-range temporal dependencies via self-attention mechanisms. Unlike recurrent architectures that build temporal understanding sequentially, transformers provide global temporal receptive fields at every layer, enabling direct modeling of relationships between distant time points (Khan et al., 2022).

3.3 Temporo-Spatial Vision Transformer (TSViT)

Tarasiou et al., 2023 introduce TSViT, the first fully-attentional model specifically designed for general SITS processing. The architecture implements a critical insight: for satellite image time series, **temporal-then-spatial factorization** proves vastly superior to the spatial-then-temporal approach used in video recognition, achieving 29.7% higher performance. This finding has profound implications for vegetation monitoring, suggesting that temporal patterns should be modeled before spatial relationships, aligning well with phenological processes where vegetation growth cycles follow temporally structured patterns.

TSViT splits SITS records into non-overlapping patches in space and time, processes spatial-location-specific time series through a temporal encoder, then aggregates temporal information into class-specific spatial feature maps processed by a spatial encoder. The problem of irregular temporal sampling is adressed by an innovative temporal positional encoding, using learned lookup tables indexed by actual acquisition dates. This improvement results in a factorized architecture design reducing computational complexity from $O(N^2)$ smaller values due to the data factorization.

3.4 Efficient Architectures: VistaFormer

While TSViT demonstrates the effectiveness of attention mechanisms for SITS, its computational requirements are still high enough to pose challenges for operational deployment on high-resolution images. MacDonald et al., 2024 address this limitation with VistaFormer, a lightweight transformer architecture achieving comparable or superior performance while requiring only 8% of TSViT's computational operations.

Among the practical innovations proposed, a multi-scale enconder-decoder architecture progressively capture hierarchical temporal patterns, while the usage of gated convolutions filter atmospheric distortions without requiring pre-trained cloud masking models. Additionally a position-free self-attention mechanism eliminates the need for complex positional encoding, and when combined with Neighbourhood Attention as an alternative to multi-head self-attention further reduces the complexity to $O(NK^2)$.

With only 1.25M parameters VistaFormer establishes that efficient transformer architectures can maintain performance while being practical for processing large satellite image archives. The lightweight decoder design using trilinear upsampling and 1D convolutions provides a path of adaptation from semantic segmentation to regression tasks such as vegetation index prediction.

3.5 Multi-Modal Extensions: ContextFormer and GreenEarthNet

While VistaFormer optimizes computational efficiency for pure imagery-based SITS analysis, Benson et al., 2024 extend transformer capabilities through multi-modal data fusion in Contextformer. Recognizing that vegetation dynamics are influenced by meteorological conditions, topography, and soil properties beyond spectral reflectance, Contextformer integrates Sentinel-2 imagery with daily meteorological observations and elevation data through a unified transformer framework.

The GreenEarthNet dataset introduced alongside Contextformer provides continental-scale infrastructure for vegetation modeling research, featuring 30 5-daily satellite images, 150 aligned meteorological observations, and rigorous temporal out-of-distribution (OOD-t) and spatial-temporal extrapolation (OOD-st) test sets. Achieving $R^2 = 0.62$ and RMSE = 0.14 on 100-day forecasts, Contextformer demonstrates that multi-modal integration substantially improves vegetation forecasting accuracy.

Compared to VistaFormer's lightweight architecture (1.25M parameters), Contextformer employs a more parameter-rich design (6.1M parameters) using Pyramid Vision Transformer (PVT) as spatial backbone. The key tradeoff lies between VistaFormer's computational efficiency for large-scale imagery processing versus Contextformer's enhanced accuracy through multi-modal data integration, establishing complementary approaches for operational vegetation monitoring depending on data availability and computational constraints.

4 Precision Agriculture Applications

4.1 Domain-Specific Considerations

Mehdipour et al., 2025 provide comprehensive analysis of Vision Transformer applications in precision agriculture, with emphasis on plant disease detection and crop monitoring. While focused primarily on single-image classification rather than time series analysis, their survey offers critical aspects of agricultural remote sensing applications:

- **Data scarcity challenges** and solutions including transfer learning from large-scale datasets (ImageNet, PlantCLEF), data augmentation strategies (including GAN-based synthetic data generation), and knowledge distillation from larger models
- **Hybrid architectures** combining CNN local feature extraction with ViT global attention mechanisms, demonstrating that agricultural applications often benefit from preserving some convolutional inductive biases for texture and spectral feature extraction
- **Interpretability requirements** for agricultural decision support, incorporating explainable AI techniques (Grad-CAM, LIME) to provide agronomists with transparent reasoning behind model predictions
- **Computational efficiency** strategies including attention pruning, model compression, and lightweight architectures necessary for deployment on edge devices in field conditions

The comprehensive analysis of 44 studies reveals that pure ViT models, hierarchical variants (Swin Transformer), and hybrid CNN-ViT architectures each demonstrate advantages depending on specific application requirements, dataset characteristics, and deployment constraints. For agricultural monitoring

tasks requiring both local spectral feature discrimination and global spatial context, hybrid approaches frequently outperform pure transformers or pure CNNs.

5 Synthesis and Conclusion

5.1 Architectural Foundations for NDVI Prediction

The reviewed literature demonstrates a clear evolution from CNN-based approaches to transformer-based architectures for satellite image time series analysis. Vision Transformers offer compelling advantages for modeling the complex spatio-temporal patterns inherent to vegetation dynamics, particularly through their ability to capture long-range dependencies, integrate multi-modal data sources, and provide global receptive fields for temporal modeling.

Analysis across domains reveals convergent insights that establish clear architectural foundations for developing vision transformer systems capable of predicting NDVI from satellite image time series:

Temporal-First Processing: The finding by Tarasiou et al., 2023 that temporal-then-spatial factorization dramatically outperforms spatial-then-temporal approaches (+29.7%) for SITS represents a fundamental design principle. For vegetation index prediction, this suggests architectures should first extract temporal features capturing phenological patterns, then aggregate spatial context, providing superior performance compared to alternative processing orders.

Irregular Temporal Sampling: The acquisition-time-specific positional encodings introduced by Tarasiou et al., 2023 address a fundamental challenge in satellite monitoring where cloud cover and orbital patterns create irregular time series. NDVI prediction architectures must explicitly handle variable temporal sampling rather than assuming regular intervals.

Multi-Modal Integration: Benson et al., 2024 demonstrate that incorporating meteorological data alongside satellite imagery substantially improves vegetation forecasting ($R^2 = 0.62$ on 100-day forecasts). NDVI prediction systems should be designed to accommodate multiple data modalities through unified transformer architectures rather than treating them as separate information streams.

Efficiency-Performance Balance: The dramatic efficiency gains demonstrated by VistaFormer and the lightweight approaches surveyed by Mehdipour et al., 2025 establish that careful architectural design can reduce computational requirements by 90% while maintaining or improving performance. For operational NDVI mapping systems processing continental-scale archives, efficiency is not optional but essential, achievable through position-free attention, factorized encoders, and lightweight decoders.

Atmospheric Noise Handling: Both VistaFormer's gated convolutions and Contextformer's cloud-aware masking demonstrate the necessity of explicitly addressing atmospheric interference within the model architecture rather than relying entirely on preprocessing, essential for robust vegetation index prediction.

5.2 Critical Research Gaps

Despite substantial progress, several critical gaps remain in applying vision transformers to vegetation index prediction:

Limited Regression-Focused Architectures: The majority of reviewed SITS transformers target classification or segmentation tasks. The reviewed work provides clear insights and optimizations for the encoder components, but dedicated architectures optimized for temporal regression rather than classification remain underexplored.

Long-Horizon Forecasting: While Contextformer demonstrates 100-day vegetation forecasting, the capability to predict NDVI for arbitrary future dates (e.g., "predict NDVI for day 250 of year") requires

further architectural development. Mechanisms for temporal extrapolation beyond training sequence lengths need investigation.

Uncertainty Quantification: Agricultural decision-making requires not just predictions but confidence estimates. Integration of uncertainty quantification mechanisms (ensemble methods, Bayesian approaches, or direct probabilistic outputs) into transformer architectures remains an open challenge.

Transfer Learning for Vegetation Indices: While transfer learning from ImageNet has proven effective for classification, optimal pre-training strategies for vegetation index regression are unclear. Should models pre-train on NDVI reconstruction tasks, related vegetation indices (EVI, LAI), or other remote sensing objectives?

Interpretability and Physical Consistency: Unlike classification tasks where attention maps provide interpretability, regression outputs require mechanisms to ensure physical plausibility (e.g., NDVI bounds, phenological consistency) and explain temporal predictions in terms of biophysical processes.

5.3 Future Directions

The architectural foundations established by TSViT, VistaFormer, and Contextformer provide clear pathways forward, with VistaFormer's lightweight encoder-decoder design offering the most direct starting point for adaptation to NDVI prediction through modification of the decoder from segmentation to temporal regression outputs. The convergence of evidence across computer vision, remote sensing, and agricultural monitoring domains establishes that vision transformer architectures represent not merely an incremental improvement but a fundamental paradigm shift in how satellite image time series can be modeled. For the specific goal of predicting NDVI maps from temporal sequences of satellite imagery, the technical foundations now exist to develop sophisticated, efficient, and operationally viable systems capable of supporting precision agriculture, vegetation monitoring, and climate change assessment applications.

Bibliographical references

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. (2023). Transformers in Remote Sensing: A Survey [Publisher: Multidisciplinary Digital Publishing Institute]. *Remote Sensing*, *15*(7), 1860. https://doi.org/10.3390/rs15071860

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision Transformers for Remote Sensing Image Classification [Publisher: Multidisciplinary Digital Publishing Institute]. *Remote Sensing*, *13*(3), 516. https://doi.org/10.3390/rs13030516

Benson, V., Robin, C., Requena-Mesa, C., Alonso, L., Carvalhais, N., Cortés, J., Gao, Z., Linscheid, N., Weynants, M., & Reichstein, M. (2024). Multi-Modal Learning for Geospatial Vegetation Forecasting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27788–27799. https://doi.org/10.1109/CVPR52733.2024.02625

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Comput. Surv.*, *54*(10s), 200:1–200:41. https://doi.org/10.1145/3505244

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

MacDonald, E., Jacoby, D., & Coady, Y. (2024). Vistaformer: Scalable vision transformers for satellite image time series segmentation. *arXiv preprint arXiv:2409.08461*.

Mehdipour, S., Mirroshandel, S. A., & Tabatabaei, S. A. (2025). Vision transformers in precision agriculture: A comprehensive survey. *arXiv preprint arXiv:2504.21706*.

Tarasiou, M., Chavez, E., & Zafeiriou, S. (2023). Vits for sits: Vision transformers for satellite image time series. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10418–10428.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions [ISSN: 2380-7504]. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 548–558. https://doi.org/10.1109/ICCV48922.2021.00061