

Multispectral Vision Transformers for Non-Destructive Quality Assessment of Fresh Produce Under Dynamic Environmental Conditions

Zhenjan Guo

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

mazhanjiang@ku.edu

Abstract

The assessment of fresh produce quality remains a central challenge in postharvest supply chains because visible appearance alone cannot reliably capture internal ripeness, firmness, moisture content, and incipient defects. Multispectral imaging provides richer spectral information, but its practical adoption is constrained by dynamic environmental conditions, hardware heterogeneity, and the need for robust interpretation. This paper presents a system-level examination of multispectral vision transformers for non-destructive quality assessment of fresh produce under variable illumination, temperature, humidity, and handling conditions. Rather than focusing on a single algorithmic innovation, the work analyzes the architectural trade-offs of transformer-based spectral-spatial representation, including global attention, hierarchical tokenization, and transfer learning, in relation to conventional convolutional and spectral processing approaches. The discussion addresses infrastructure requirements, data governance, dataset bias, calibration drift, edge deployment, and sustainability. It considers how self-attention can integrate spectral and spatial context without requiring explicit feature engineering, while also noting computational costs and the need for curated spectral datasets. The paper further examines robustness mechanisms including spectral normalization, domain adaptation, and uncertainty-aware decision making. Cross-domain lessons from remote sensing, biomedical imaging, and precision agriculture are used to contextualize design choices. The analysis suggests that vision transformers can serve as an enabling backbone for resilient quality assessment systems, provided that institutional and technical safeguards are aligned with operational realities. A coordinated approach spanning sensing, model governance, and supply chain integration is necessary to translate laboratory performance into dependable field deployment.

Keywords

multispectral imaging; vision transformers; fresh produce quality; non-destructive assessment; dynamic environments; system architecture.

1. Introduction

Advances in computer vision have entered a period of architectural transformation through vision transformers, which model images as sequences of patches and apply self-attention rather than relying solely on convolutional inductive biases [1]. This development extends a broader shift initiated by attention-based sequence modeling, which demonstrated that global dependencies can be learned directly from structured representations without recurrent processing [2]. In parallel, hyperspectral and multispectral imaging have become established in food and agricultural quality inspection because they capture chemical and structural

properties beyond the visible range, enabling the detection of bruises, sugar content gradients, water stress, and early decay that would otherwise remain hidden from conventional color cameras [3], [4]. These two trajectories create an opportunity to rethink the design of non-destructive quality assessment systems, moving from handcrafted spectral indices toward learned spectral-spatial representations that can adapt to complex operational settings.

Traditional chemical and sensory methods for assessing fresh produce are destructive, slow, and unsuitable for high-throughput sorting or continuous supply chain monitoring. Near-infrared spectroscopy has long demonstrated that internal quality attributes such as soluble solids content, acidity, and dry matter can be estimated without damaging the product, but point-based spectroscopic measurements often fail to capture spatial heterogeneity across the surface of a fruit or vegetable [5]. Deep convolutional networks improved spatial feature extraction in quality inspection tasks by learning hierarchical visual patterns from large labeled datasets [6], while encoder-decoder segmentation architectures extended this capability to dense pixel-level prediction in biological and medical imaging contexts [7]. Object detection transformers subsequently showed that attention-based sequence modeling can be applied directly to detection tasks without the need for manually designed region proposals [8]. Hierarchical vision transformers further addressed the computational and representational limitations of early patch-based transformers by introducing shifted window attention and multiscale token merging [9]. Recent work on compact multispectral imaging and machine learning systems has shown the feasibility of real-time dynamic reflectance monitoring in settings where illumination and object motion are not strictly controlled [10]. Despite these advances, the deployment of such technologies in fresh produce supply chains remains constrained by system-level challenges rather than by a single architectural choice.

This paper examines multispectral vision transformers as part of a larger socio-technical infrastructure for fresh produce quality assessment. The focus is not on incremental benchmark improvements but on the structural trade-offs among global contextual modeling, spectral calibration, computational cost, data governance, and field robustness. The following sections discuss spectral sensing foundations, transformer architectures, environmental variability, deployment architectures, governance, and sustainability. The analysis integrates lessons from precision agriculture, remote sensing, and biomedical photonics to provide a cross-domain perspective on how attention-based spectral-spatial models can be responsibly integrated into dynamic postharvest environments.

2. Multispectral Sensing and Quality Assessment Foundations

Multispectral imaging occupies an intermediate position between conventional red-green-blue imaging and dense hyperspectral acquisition. It typically captures a limited number of spectral bands selected according to the absorption features of water, chlorophyll, carotenoids, and other constituents relevant to fresh produce quality [3], [4]. Compared with hyperspectral imaging, multispectral systems can be less expensive, more mechanically robust, and easier to operate at high throughput, but they require careful band selection because the discriminative information must be encoded in a reduced spectral space. The biochemical basis of quality assessment relies on the interaction of light with tissue structure and chemical compounds, producing reflectance, absorbance, and scattering signatures that vary with ripeness, firmness, and internal damage [5]. In practical systems, these signatures are also influenced by surface curvature, skin texture, moisture, and the angular geometry of illumination and detection.

However, building predictive models from spectral imagery is hindered by limited labeled produce batches, seasonal variability, and cultivar differences. The development of data-

efficient vision transformer training strategies is therefore directly relevant because it suggests that strong representations can be learned from smaller datasets when knowledge is distilled from larger teacher models and appropriate data augmentation is applied [11]. Machine learning has been widely applied in precision agriculture for crop yield prediction and nitrogen status estimation [12], but adaptation to postharvest quality variables requires recalibration because the target properties, measurement geometries, and biological dynamics differ substantially. Survey evidence indicates that deep learning adoption in agriculture has accelerated across plant recognition, disease detection, and yield estimation, yet many applications remain constrained by dataset scarcity and limited generalization across environments [13]. Spectral transformer architectures have extended self-attention to hyperspectral image classification by treating spectral channels as trainable representations rather than fixed input dimensions [14]. Reviews of deep learning classifiers for hyperspectral data similarly emphasize that global context modeling can capture subtle spectral-spatial relationships that are difficult to encode through handcrafted features [15].

From a system design perspective, the selection of spectral bands and model architecture cannot be separated from the downstream decision context. A sorting line operating at several fruits per second requires low-latency inference and stable illumination, whereas a handheld field device may need to tolerate variable sunlight and limited battery capacity. These constraints influence whether the system should use a global transformer with raw spectral tokens, a hierarchical transformer with spatial downsampling, or a hybrid architecture that uses convolutional layers for local spectral compression before attention-based global reasoning. The appropriate balance shifts when the system must also support traceability, explainability, and recalibration across batches, because different architectural choices expose different interfaces for inspection and governance.

3. Vision Transformer Architectures for Spectral-Spatial Representation

Vision transformers convert an image into a sequence of patches, embed each patch into a vector space, and then apply multiple layers of multi-head self-attention to capture long-range dependencies [1], [2]. In multispectral applications, each token can represent a spatial patch across all spectral bands, or separate tokens can be constructed for spectral bands and spatial locations. The first approach preserves the native spectral-spatial alignment and allows attention to learn correlations between distant regions and spectral features. The second approach increases token count and can improve spectral disentanglement but introduces additional positional and modal alignment challenges. Global self-attention provides a powerful mechanism for integrating context from the entire fruit surface, which is valuable when defects are subtle, diffuse, or distributed across multiple locations.

The main architectural trade-off arises from the quadratic computational cost of global attention with respect to token count. Hierarchical vision transformers reduce this cost by computing attention within local windows and progressively merging tokens to form multiscale representations [9]. For multispectral images with moderate spatial resolution, hierarchical designs may be preferable for edge deployment because they limit memory consumption while retaining the ability to model broad context at later stages. In contrast, early global attention may be more appropriate for small regions of interest where the entire produce item can be represented by a modest number of tokens. Data-efficient training techniques further mitigate the need for enormous labeled datasets by combining patch-level distillation with strong augmentation and careful regularization [11]. Convolutional encoders and segmentation networks remain useful as spectral compressors or as hybrid components

that transform raw multispectral cubes into compact token sequences before transformer layers [6], [7]. Detection-oriented transformers illustrate that set-based prediction frameworks can be adapted to produce instance-level outputs for items on a conveyor belt, thereby connecting spectral representation with sorting decisions [8].

Spectral transformer designs are particularly relevant because they treat spectral signatures as learnable vectors, enabling the model to suppress noisy bands and amplify diagnostically useful absorption features [14]. Reviews of deep hyperspectral classification emphasize that convolutional and recurrent models have been strong baselines, but transformer-based methods can outperform them when sufficient training data and spectral diversity are available [15]. In fresh produce applications, spectral diversity is often limited by cultivar and season, so transfer from larger agricultural or remote sensing datasets may be required. The compact multispectral monitoring systems described in recent work demonstrate that machine learning can process dynamic reflectance in real time, but they also highlight the importance of stable optical calibration and model adaptation when environmental conditions shift [10]. The architectural choice is therefore not merely a technical detail but a governance decision that affects how easily the system can be inspected, recalibrated, and certified.

4. Dynamic Environmental Variability and Robustness Trade-offs

Fresh produce quality assessment in the field or in open packing facilities operates under conditions that differ substantially from controlled laboratory imaging. Illumination intensity and spectral composition change throughout the day, while temperature and humidity alter both the optical properties of produce tissue and the response of imaging sensors. Internet of Things platforms in agriculture have enabled continuous monitoring of microclimatic variables, but they also generate heterogeneous data streams that must be synchronized with spectral images [16]. Big data research in smart farming emphasizes that data variety, velocity, and veracity are often more problematic than data volume, because missing calibration metadata and inconsistent sensor placement can degrade model performance despite large sample counts [17]. These challenges are intensified in postharvest environments where produce moves rapidly, rotates, and may be partially obscured by leaves, stems, or packaging.

Robustness to environmental variability can be pursued through several complementary strategies. Spectral normalization attempts to remove illumination-dependent changes by referencing standard surfaces or estimating the incident spectrum from image statistics. Domain adaptation methods align feature distributions across seasons, cultivars, or sensor units without requiring full recalibration. Uncertainty-aware decision making introduces abstention mechanisms that flag ambiguous predictions for manual inspection when the model encounters conditions outside its training distribution. Plant disease identification studies have shown that deep learning models trained under specific imaging conditions often fail when exposed to new lighting, backgrounds, or symptomatic variations, indicating that robustness cannot be assumed from high internal validation accuracy [18]. Close-range hyperspectral imaging of plants similarly demonstrates that leaf orientation, moisture, and canopy structure produce complex bidirectional reflectance effects that must be explicitly managed [19]. Cross-domain experience from geological and environmental remote sensing reinforces the lesson that spectral unmixing and atmospheric correction are essential before machine learning models are applied to spatially extended scenes [20].

Vision transformers can support these robustness strategies through their capacity to model global context, but they also introduce new failure modes. Attention may learn shortcuts based on illumination gradients or sensor artifacts if the training distribution is not carefully

curated. A model that appears accurate on average may nevertheless produce systematically biased predictions for darker produce, wet surfaces, or damaged regions where spectral reflectance deviates from the learned manifold. Therefore, robustness monitoring must be treated as an ongoing system process rather than a one-time validation exercise. This includes logging environmental covariates, tracking prediction confidence, and periodically auditing model outputs against destructive reference measurements.

5. System Integration and Deployment Architectures

Deploying multispectral vision transformers in operational fresh produce supply chains requires integrating imaging hardware, illumination control, edge computing, cloud storage, and decision support workflows. A central design tension exists between the computational demands of transformer inference and the latency constraints of high-speed sorting. Cloud-based processing enables flexible model updates and centralized quality monitoring, but it introduces communication latency, privacy concerns, and dependence on network availability. Edge deployment reduces latency and can operate in facilities with limited connectivity, but it restricts model size and complicates the distribution of software updates. Hybrid architectures can balance these demands by performing lightweight spectral calibration and region proposal on edge devices while reserving cloud resources for periodic retraining and fleet-level analytics.

The choice of deployment architecture also affects data governance. A centralized system can accumulate large multispectral datasets that support continuous model improvement and cross-cultivar generalization, but it may concentrate control over sensitive business information such as harvest volumes, supplier quality, and defect rates. A distributed system can preserve local autonomy and reduce data transfer costs, yet it may produce fragmented datasets that are insufficient for training globally robust models. Systematic reviews of machine learning for crop yield prediction illustrate that data management practices, feature consistency, and model updating procedures are as important as algorithmic selection for long-term performance [21]. In environmental remote sensing, deep learning systems have demonstrated that multi-source data fusion and modular preprocessing pipelines are essential for managing diverse acquisition conditions and maintaining transferability across regions [22].

Multispectral vision transformer systems must also address calibration drift over time. Optical components age, illumination sources shift, and produce cultivars change across growing seasons. A deployable system should include scheduled reference measurements, automated spectral quality checks, and versioned model artifacts that can be traced to specific training datasets and calibration procedures. The system should be designed so that a new model can be rolled back without disrupting sorting operations if unexpected performance degradation is detected. This requires treating machine learning models as maintainable software components embedded within a broader sensing and decision infrastructure, rather than as standalone analytical tools.

6. Governance, Sustainability, and Policy Implications

The integration of multispectral vision transformers into fresh produce supply chains raises governance questions that extend beyond technical accuracy. Quality assessment models can influence market access, supplier payments, waste flows, and food safety determinations. If a model systematically misclassifies certain cultivars, regions, or visual appearances, it may produce unfair economic outcomes for growers and distributors. Responsible deployment

therefore requires transparency about training data composition, model limitations, and the conditions under which automated decisions can be overridden by human inspectors. Auditability is especially important in high-stakes contexts where non-destructive predictions about ripeness or shelf life affect contractual obligations.

Sustainability implications are also significant. Non-destructive quality assessment has the potential to reduce food loss by enabling better sorting, routing, and storage decisions, but the environmental footprint of the digital infrastructure must be considered. Transformer models can be computationally intensive during training, and large-scale data collection may require energy-consuming illumination, cooling, and data storage systems. Hierarchical architectures, efficient attention mechanisms, and selective data acquisition can reduce computational burdens while maintaining adequate accuracy [9], [11]. Designing systems with modular hardware and open data formats can extend operational lifetimes and reduce electronic waste.

Policy frameworks should encourage interoperability, data sharing, and standardization of spectral calibration without imposing rigid prescriptions that stifle innovation. Public research institutions can play a role in creating benchmark datasets that span multiple cultivars, seasons, and environmental conditions, thereby reducing the advantage of large private data holders and improving scientific reproducibility. Certification procedures for non-destructive quality assessment systems may need to include explicit criteria for robustness against dynamic environmental conditions, calibrated uncertainty reporting, and nondiscrimination across produce categories. These governance measures should be proportionate to risk, allowing low-stakes internal quality inspection to proceed with lighter oversight while applying stricter scrutiny to systems that determine market access or contractual quality grades.

7. Conclusion

Multispectral vision transformers represent a promising convergence of spectral sensing and attention-based representation learning for non-destructive quality assessment of fresh produce. Their capacity to integrate spectral and spatial context can improve detection of subtle internal and external defects, but operational success depends on more than architectural innovation. Robust deployment requires careful attention to spectral calibration, environmental variability, computational constraints, data governance, and institutional accountability. The system-level perspective presented in this paper highlights the structural trade-offs among global attention, hierarchical tokenization, edge deployment, calibration maintenance, and supply chain integration. These trade-offs cannot be resolved by any single technical choice; they require coordinated design across sensing hardware, model development, data infrastructure, and quality management policy. Future work should investigate longitudinal field evaluations that expose models to multiple seasons and cultivars, standardized benchmarks that include realistic environmental perturbations, and governance protocols that enable transparent auditing without imposing unsustainable compliance burdens. By treating multispectral vision transformers as components of a broader socio-technical system, researchers and practitioners can move beyond laboratory demonstrations toward dependable, fair, and sustainable quality assessment infrastructures.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021).

An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
3. Lorente, D., Aleixos, N., Gómez-Sanchis, J., Cubero, S., García-Navarrete, O. L., & Blasco, J. (2012). Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food and Bioprocess Technology*, 5(4), 1121–1142.
4. Qin, J., Chao, K., Kim, M. S., Lu, R., & Burks, T. F. (2013). Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, 118(2), 157–171.
5. Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, 46(2), 99–118.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer.
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213–229). Springer.
9. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).
10. Tang, G., Zheng, I., Teal, P. D., Zhang, F. L., & Doronin, A. (2026, March). Real-time monitoring of dynamic reflectance using a compact multispectral imaging and machine learning system. In *Dynamics and Fluctuations in Biomedical Photonics XXIII* (Vol. 13850, pp. 4-11). SPIE.
11. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning* (pp. 10347–10357). PMLR.
12. Chlingaryan, A., Sukkariéh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69.
13. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.
14. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., & Chanussot, J. (2022). SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

15. Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 279–317.
16. Tzounis, A., Katsoulas, N., Bartzanas, T., & Kittas, C. (2017). Internet of Things in agriculture, recent advances and future challenges. *Biosystems Engineering*, 164, 31–48.
17. Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming—a review. *Agricultural Systems*, 153, 69–80.
18. Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, 180, 96–107.
19. Mishra, P., Asaari, M. S. M., Herrero-Langreo, A., Lohumi, S., Diezma, B., & Scheunders, P. (2017). Close range hyperspectral imaging of plants: A review. *Biosystems Engineering*, 164, 49–67.
20. Van der Meer, F. D., van der Werff, H. M. A., van Ruitenbeek, F. J. A., Hecker, C. A., Bakker, W. H., Noomen, M. F., van der Meijde, M., Carranza, E. J. M., de Smeth, J. B., & Woldai, T. (2012). Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 14(1), 112–128.
21. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709.
22. Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241, 111716.