Journal of Animal & Plant Sciences, 36(2): 2026, Page: 321-330
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2026.2.0027
Review Article

A COMPREHENSIVE REVIEW OF OBJECT DETECTION IN ANIMAL AND PLANT USING VISION TRANSFORMER

M. Lin[1]·Z. Gao[2, 3]·W. Liao[1*] and H. Cai[1*]

[1.]Department of Physics, School of Science, Jimei University, Xiamen, Fujian Province, China
[2.]School of Information Science and Technology, Shanghai Tech University, Shanghai, China
[3.]Guangdong Institution of Intelligent Science and Technology, Zhuhai, Guangdong Province, China
[*]Corresponding authors' email: Honghao Cai, hhcai@jmu.edu.cn; Wenliang Liao, 200661000118@jmu.edu.cn
ORCIDs: Honghao Cai: https://orcid.org/0000-0002-1870-8061

ABSTRACT

In digital farming, computers serve as primary sensing eyes and object detection is the core vision task that locates and counts the target objects of interest, i.e., plants, fruits and livestock, in various agricultural systems. While, Vision Transformers (ViTs), a natural language processing alternative to convolutional neural networks by capturing global context through self-attention, have shown great potential in object detection. However, the field of ViT-based detectors remains fragmented, with independent advances in plant and animal studies and a lack of comprehensive analysis connecting these domains. To bridge this gap, we conducted a systematic review, retaining 30 primary studies after a dual screening and quality appraisal process—20 focused on plant production and 10 on animal production. Our analysis shows that ViT-based models excel in multi-scale representation, complex scene reasoning, and efficient feature extraction. These capabilities give high accuracy in fruit quality assessment, crop growth monitoring, weed detection, meat grading and livestock behaviour surveillance. However, challenges such as high computational complexity, large parameter sizes, environmental variability, small object detection, and data annotation requirements remain. For researchers and practitioners, this review offers a unified framework to understand ViT-based detection. It pinpoints cross-domain challenges and concludes with a forward-looking pathway to turn these insights into practical, on-farm solutions.

Keywords: Convolutional neural network; Computer vision; Deep learning; Machine learning; ViT; YOLO

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/

Published first online December 25, 2025 Published final February 28, 2026

INTRODUCTION

Plant science and animal science increasingly rely on high-resolution vision systems to phenotype crops and monitor livestock at the organismal or sub-organismal level. Accurate object detection—whether counting apples in an orchard canopy or localising claw lesions on a dairy cow—directly informs breeding decisions, health interventions, and resource allocation. Traditional convolutional networks have served these disciplines well (Lu *et al.*, 2021; Kumar *et al.*, 2023; Gao *et al.*, 2024a; Liu *et al.*, 2025a), but their receptive-field limitations and sensitivity to domain shift frequently translate into missed detections and costly false positives under unconstrained field or barn conditions (Pei *et al.*, 2025; Sui *et al.*, 2025). The Vision Transformer (ViT) has recently challenged this status quo by replacing local convolution with global self-attention, enabling a single model to capture long-range interactions across entire images (Han *et al.*, 2022). Early plant-science reports show improved lesion delineation (Singh *et al.*, 2024), and parallel livestock studies demonstrate cow behavior recognition (Geng *et al.*, 2024), yet no systematic effort has synthesised these plant- and animal-centric advances or examined how ViT's unique strengths and computational demands translate across the two domains.

Despite isolated success stories, the plant- and animal-science communities have pursued ViT-based detection largely in parallel, using disparate datasets, evaluation metrics, and experimental protocols. Plant researchers typically benchmark on greenhouse or UAV imagery with densely annotated leaves, fruits, and diseases (Jamali *et al.*, 2025; Zhang *et al.*, 2025), whereas animal scientists focus on RGB or depth sequences acquired in barn aisles and feeding alleys to pinpoint individual cattle or lesions on livestock (Xu *et al.*, 2024; Li *et al.*, 2025). This disciplinary divide has produced a fragmented evidence base: accuracy gains are reported relative to different Convolutional Neural Network (CNN) baselines, compute budgets are quoted without standardised hardware, and claims of robustness to illumination, occlusion, or sensor variation remain

anecdotal. Consequently, breeders, veterinarians, and precision-agriculture engineers lack actionable guidance on which ViT variant to adopt, how much annotated data are truly necessary, and whether a single model can generalise from greenhouse seedlings to pasture cows.

To close this gap, we present the first systematic review dedicated to Vision Transformer–based object detection across both plant and animal sciences. We surveyed 223 peer-reviewed and preprint articles published between 2017 and August 2025, extracted 30 studies that met stringent inclusion criteria, and stratified findings by domain (20 plant-focused, 10 animal-focused). This review provides plant and animal scientists with a common reference point, accelerates technology transfer between the two communities, and charts a research agenda toward robust, resource-efficient ViT systems for organism-level phenotyping and health monitoring.

Transformer Architecture

From Language to Vision: Originally introduced for machine translation (Vaswani, 2017), the Transformer replaces recurrence with self-attention, enabling parallel training and long-range dependency modelling. Vision Transformer adapts this idea to images by treating each non-overlapping 16 × 16 pixel patch as a "token", linearly embedding it, and feeding the sequence into a stack of identical Transformer blocks (Dosovitskiy *et al.*, 2020). Positional encodings supply spatial order, while the final classification token (or per-patch tokens in detection) produces task-specific outputs.

Core Components: The Vision Transformer architecture (As shown in Figure 1) effectively processes image patches by integrating several key components. As a core component, the Multi-Head Self-Attention (MSA) mechanism globally computes query-key-value interactions among all patches within a single layer to capture contextual relationships. Each MSA module is followed by a Feed-Forward Network for non-linear feature transformation, which typically consists of two linear layers with a Gaussian Error Linear Unit activation and is often enhanced by squeeze-and-excitation layers for channel-wise compression. To ensure stable training and enable the construction of deeper models, each of these sub-layers is equipped with Layer Normalisation and Residual Connections. Finally, spatial information is incorporated through Positional Encoding, which employs either absolute sinusoidal or learnable vectors, while modern variants like the Swin Transformer often adopt a more efficient relative position bias instead.

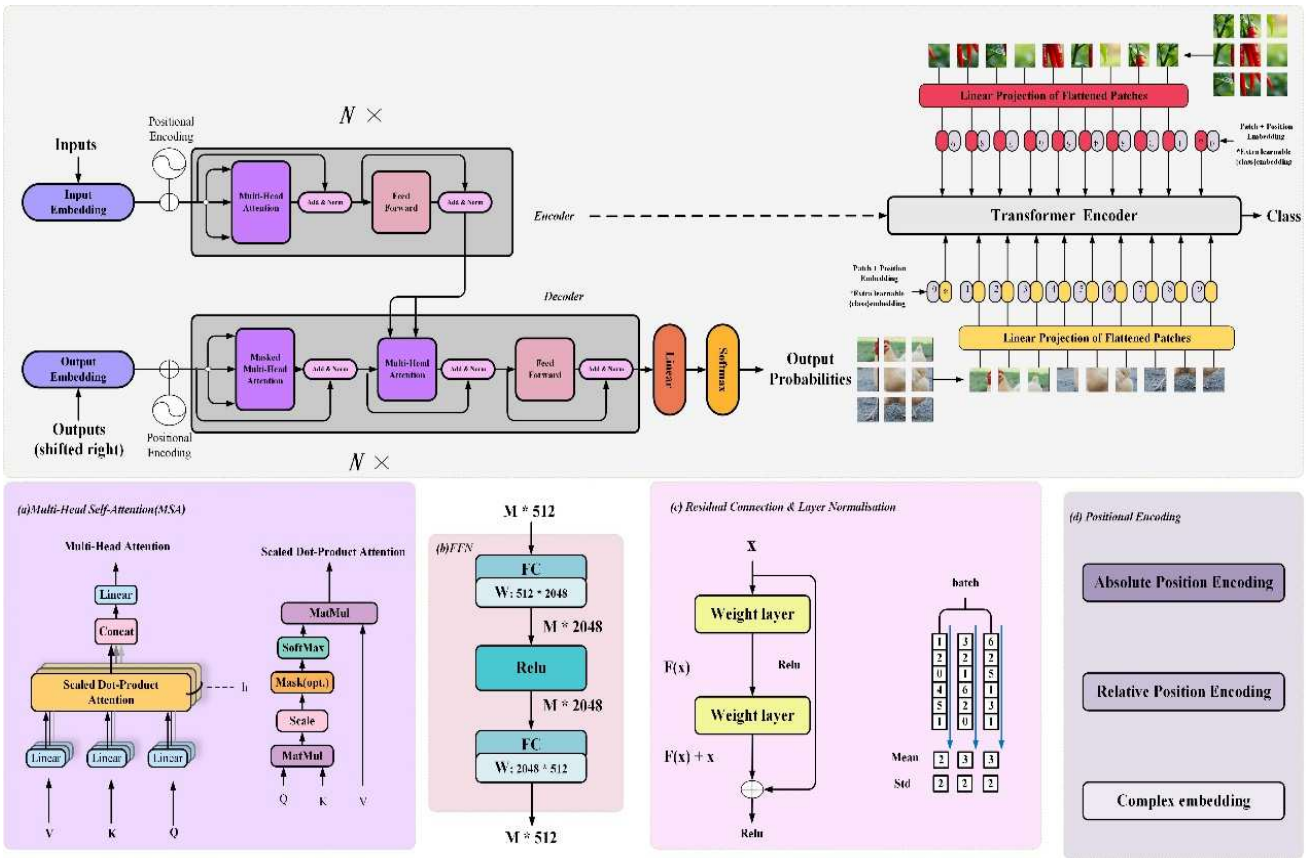

Figure 1. Transformer architecture and main components.

Variants of ViTs for Detection

Hierarchical Pyramid

Pyramid Vision Transformer (PVT): Differing from the original Vision Transformer which outputs feature maps of single scale, the PVT designs a hierarchical pyramid structure (Wang *et al.*, 2021). Such hierarchical pyramid design is essential for subsequent dense prediction tasks, such as object detection, since PVT could obtain multi-scale feature maps. Such characteristic makes PVT a suitable backbone to identify objects of various sizes ranging from small fruits to large livestock.

Swin Transformer: Due to computing self-attention over the entire image, the computational cost is expensive for dense prediction tasks. To alleviate this problem, the Swin Transformer (Liu *et al.*, 2021) computes self-attention in non-overlapping local windows. To allow connection among cross-window self-attention, the window partitions are shifted in the successive layer. Such design makes Swin Transformer a suitable backbone to identify objects of various sizes in complex scenes of agricultural setting.

Twins: The Twins (Chu *et al.*, 2021) addresses the main limitation of pure transformers: their lack of spatial locality. It combines the fine-grained local features extracted by convolution and long-range global dependencies extracted by self-attention. Such design proves to be effective for accurately detecting occluded or densely packed objects in agricultural setting by harnessing both fine-grained local features and long-range global dependencies.

Query-Based Detectors

Detection Transformer (DETR): A paradigm shift in detection architecture was introduced by the DETR (Carion *et al.*, 2020), which frames object detection as a set prediction problem, eliminating the need for hand-designed components like anchor boxes and non-maximum suppression.

Deformable DETR: To address this slow convergence, Zhu et al. proposed Deformable DETR (Zhu *et al.*, 2020), which employs a deformable attention module to allocate computation resources to only a small set of key sampling points around a reference. This not only improves efficiency but also enables more accurate predictions in more complex scenes with occlusion, which is frequently encountered in agricultural applications.

Hybrid CNN-ViT

Lightweight Vision Transformer (LeViT): For real-time agricultural applications on edge devices, model efficiency is paramount. LeViT (Graham *et al.*, 2021) tackles this by integrating convolutional layers into the transformer backbone to distill spatial features more efficiently before applying attention. The resulting hybrid model has the representational power of ViTs while achieving a desirable accuracy ‑ speed trade-off, making it feasible to deploy on-farm.

MobileViT: Targeting resource-constrained deployment, MobileViT (Mehta *et al.*, 2021) introduces a lightweight hybrid architecture that reinterprets the self-attention mechanism of Transformers as a convolutional operation. This design allows the model to capture global contextual information like a Transformer while maintaining the spatial bias and efficiency of a CNN, making it particularly effective for mobile and edge devices.

FasterViT: Instead of applying generic acceleration techniques on the Vision Transformer, FasterViT's (Hatamizadeh *et al.*, 2023) innovative contribution is to employ a hierarchical architecture in which efficient convolutional stages are used in earlier layers, and a novel window-based attention mechanism is applied in deeper layers to balance the computational load. This enables much faster inference at an acceptable accuracy drop.

Vision Transformer in Animal and Plant

Vision Transformers and their latest variants are rapidly becoming standard tools in plant and animal sciences. As illustrated in Figure 2, in plant science, published peer-reviewed papers already report use in (i) fruit physiological-quality evaluation, (ii) in-field crop-growth monitoring and stress detection, (iii) early plant-disease diagnosis at leaf and canopy levels, and (iv) site-specific weed discrimination for precision spraying. In animal science, recent work shows ViT-based models being superior at (v) automated grading of meat and other animal products and (vi) continuous livestock behaviour and health monitoring. Table 1 collates the core information of the representative studies. It not only concisely presents the current research landscape of this field but also offers a reference framework for subsequent analyses of the technical characteristics and application bottlenecks in this direction.

Fruit Physiological-quality Evaluation: Multiple studies demonstrate ViT's superiority over conventional CNN approaches for various fruit evaluation tasks. (Shimazu *et al.*, 2025) applied a ViT model to estimate the color and sensory evaluation of Shine Muscat grapes from standard camera images. Similarly, (Da Silva Ferreira *et al.*, 2024) found that ViT Small architecture provided superior classification capabilities for post-harvest dragon fruit assessment compared to both ViT Tiny and ResNet alternatives. The adaptability of ViT frameworks extends across diverse fruit types, as evidenced by (Apostolopoulos *et al.*, 2023), who presented a comprehensive machine learning framework

based on ViT to evaluate the quality of a general range of fruits. Their approach leveraged deep image features extracted from high-resolution fruit imagery, demonstrating the versatility of transformer-based architectures in agricultural quality control applications.

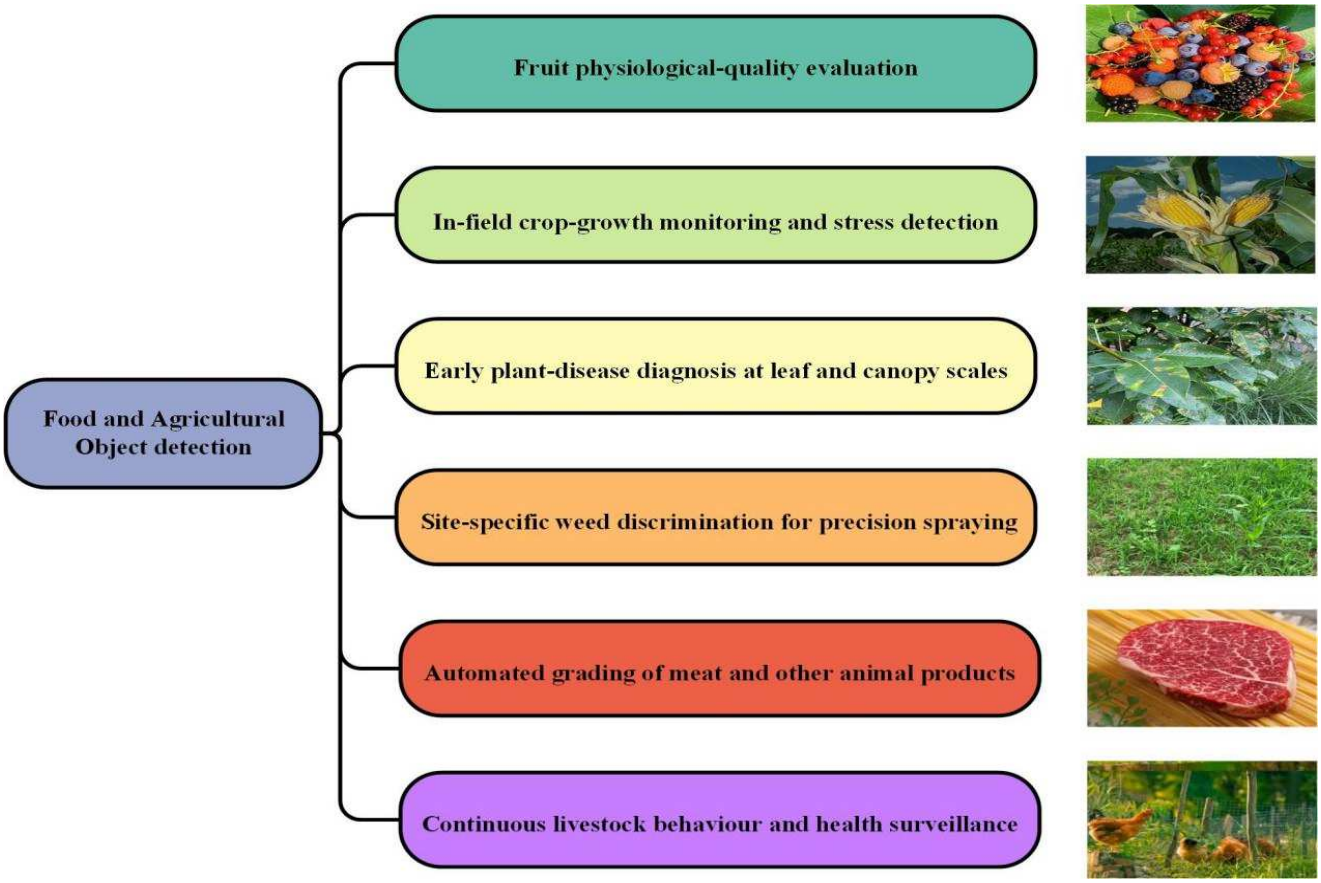

Figure 2. Common applications of visual Transformer in animal and plant object detection

In-field Crop-growth Monitoring and Stress Detection: For cereal crops, (Ni *et al.*, 2024) developed a hybrid architecture combining convolutional operations with self-attention mechanisms specifically designed for deployment on UAV platforms. Their Maize Hybrid Vision Transformer effectively balances computational efficiency with high performance for both growth stage monitoring and nitrogen stress detection. This approach represents an important advancement in lightweight transformer models suitable for edge deployment in precision agriculture applications. In horticultural systems, water stress detection has been enhanced through transformer-based approaches that can effectively capture the spatial relationships in plant imagery. (Koike *et al.*, 2024) demonstrated how ViT models can leverage their inherent ability to model long-range dependencies to assess water stress in tomato plants through leaf wilting patterns. These studies collectively highlight how transformer-based architectures are overcoming traditional limitations in field-based crop monitoring.

Site-Specific Weed Discrimination for Precision Spraying: Recent research has focused on developing hybrid models that combine the strengths of convolutional neural networks with the contextual awareness of transformer mechanisms to address specific challenges in crop protection and quality management. For disease management in vineyards, (Fu *et al.*, 2024) proposed a hybrid model called the Ghost-convolution-enlightened Transformer (GhostViT), which integrates a Ghost Module-inspired ViT encoder for detecting grape leaf diseases and insect pests. The Ghost Module enables the Transformer model to have the ability to extract multi-level features from images and decrease the parameter amount of the model and reduce the computational cost without compromising the accuracy on the task. In post-harvest quality assessment, (Dümen *et al.*, 2024) conducted a comprehensive evaluation of transformer architectures against traditional CNN approaches for citrus quality classification. The study's rigorous comparative methodology, including extensive data augmentation strategies, establishes an important benchmark for future research in this domain.

Table 1：Vision Transformer applications in agricultural and animal sciences

Research Field	Main Methods	Key Metrics	Results
Fruit Physiological-Quality Evaluation - Shine Muscat Grapes	ViT	Color estimation accuracy	97.2% (outperforms conventional CNN models)
Fruit Physiological-Quality Evaluation - Dragon Fruit	ViT Small model (compared with ViT Tiny and ResNet)	Overall classification accuracy	91.0%
General Fruit Quality Assessment - Apples, Grapes, Tomatoes	ViT-based comprehensive machine learning framework (trained on deep features from high-resolution images)	Quality classification accuracy	Apples: 99.50%; Grapes: 100%; Tomatoes: 99.50%
In-field Crop-Growth Monitoring – Maize	Maize Hybrid Vision Transformer (integrating convolutional and self-attention mechanisms, deployed on UAVs)	Growth monitoring/nitrogen stress detection accuracy, model parameters	224×224 input: 97.71% accuracy; 512×512 input: 98.71% accuracy; 15.446 million parameters
In-field Crop Stress Detection – Tomato	ViT model (combined with unsupervised domain adaptation)	Correlation coefficient for water stress estimation	0.82 (correlated with relative stem diameter, RSD)
Crop Disease/Insect Pest Detection - Grape Leaves	Ghost-convolution-enlightened Transformer (GhostViT, integrating Ghost Module-inspired ViT encoder)	Disease/insect pest detection accuracy	98.14% (reduces parameters and computational cost without accuracy loss)
Fruit Quality Classification – Lemons	ViT, Swin Transformer, and 8 CNN models (Xception, ResNet-50, InceptionV3, etc., with data augmentation)	Quality classification accuracy	ViT: 99.84% (best); ResNet-50: 99.78%; InceptionV3: 99.69%; Swin Transformer: 99.23%
Site-Specific Weed Discrimination - Soybean Fields	YOLO-SW model (Swin Transformer backbone + RT-DETR encoder + CARAFE dynamic upsampling)	Weed detection mAP	mAP@50: 92.3% (outperforms state-of-the-art methods)
Precision Spraying Pest Detection - Cabbage Farming	Modified YOLOv5 model (added Transformer module, fusing CNN and ViT advantages)	Target detection mAP	94.2% (outperforms standard YOLOv5)
Automated Meat Grading - Mutton Multi-part Classification	Swin Transformer model	Cut classification accuracy	Higher than traditional CNNs and other deep learning models; enables precise/efficient meat classification
Meat Adulteration Detection - Beef Carrageenan	Swin-Tiny (Swin-T) model (smartphone-captured images)	Top-1 accuracy, detection speed, model size	Top-1 accuracy: 99.7%; Detection speed: 3.2 ms; Model size: 103.45 Mb (outperforms EIS method's 79.2%)
Livestock Health Surveillance - Cow Tracking & Feeding Monitoring	YOLOv5s-CA+DeepSORT-ViT hybrid framework (added Coordinate Attention (CA); ViT replaces DeepSORT sorter)	F1 score, multi-target tracking accuracy, processing time	F1 score: 88.5%; Multi-target tracking accuracy: 84.4%; Significantly reduced processing time
Livestock Re-identification - Chicken	Transformer-based model (adopting similarity learning strategy	Top-1 accuracy (small/large flocks)	Small flocks: >80%; Large flocks (100 chickens): ~40% (scalability challenges remain)

For field-based weed management, researchers have developed specialized architectures that address the unique challenges of high-resolution aerial imagery analysis, (Shuai *et al.*, 2025) enhanced detection capabilities in soybean fields by creating a hybrid model that integrates Swin Transformer backbones with real-time detection mechanisms. Similarly, (Fu *et al.*, 2022)

demonstrated how transformer modules can be effectively integrated with established CNN frameworks like YOLOv5 to improve contextual understanding in cabbage farming applications. Both studies highlight the importance of capturing long-range dependencies in agricultural imagery for precise intervention targeting.

Automated grading of meat and other animal products: Transformer-based architectures are revolutionizing quality assessment and authentication protocols in meat processing and food safety applications. In meat processing facilities, automated classification systems are increasingly replacing manual inspection methods. (Zhao *et al.*, 2023) used the Swin Transformer to identify and classify different cuts of mutton automatically. Food safety applications also have particularly benefited from transformer-based approaches. (Gao *et al.*, 2024b) pioneered a smartphone-based detection system for identifying hydrocolloid adulteration in beef products. Their implementation of Swin Transformer models demonstrates how advanced computer vision can be deployed on accessible hardware to address critical food safety challenges.

Continuous Livestock Behaviour and Health Surveillance: (Guo *et al.*, 2023) aim to address the challenges of tracking cows and monitoring feeding behaviors in commercial farms by designing a hybrid framework, YOLOv5s-CA+DeepSORT-ViT. To be more specific, they apply a Coordinate Attention (CA) mechanism to enhance the sensitivity of YOLOv5s to positional information and further replace the sorter in DeepSORT with a Vision Transformer to enhance global feature matching, which effectively reduces identity switches. This work demonstrates a practical integration of advanced attention mechanisms and transformers to optimize real-world livestock management workflows. Poultry monitoring presents distinct challenges due to the visual similarity between individual birds and the density of commercial housing systems. (Lamping *et al.*, 2025) investigated the possibility of re-identifying individual hens in uncontrolled settings with a transformer-based model. The authors' innovation was a similarity-learning strategy that obviated the need for per-chicken training images. When comparing to the authors' Re-id model, it was found that Top-1 accuracy was greater than 80% for small groups of hens, but only ~40% for larger flocks of 100 hens. Thus, there is potential for scalability issues. Nonetheless, the model learned to focus on biologically salient features (combs and wattles) much like those that a human would use in identification. This work represents the first step towards demonstrating that re-identification with transformers is possible in commercial poultry farming, but more work is needed to increase accuracy at scale.

Challenges and Future in Plant and Animal Sciences: Despite the promising potential and current applications, there are still many challenges that need to be addressed in plant and animal sciences.

High Computation Consumption in Agricultural and Veterinary Applications: The self-attention mechanism in Vision Transformers, which computes pairwise interactions between all image patches, suffers from quadratic complexity relative to input size (Zhang *et al.*, 2024). This computational cost is a roadblock to the use of ViTs in plant and animal sciences, where high resolution images are often necessary in order to capture fine-grained distinctions—like in the case of identifying weeds in real-time while operating a combine harvester, or monitoring animal health in a feed lot—thus often requiring significant computational resources. Several strategies to alleviate this shortcoming have been proposed. One approach is to use sparse attention mechanisms (Zhao *et al.*, 2024; He *et al.*, 2025), which limit the attention span of a patch to a small subset of the most relevant patches, thereby reducing computational cost. Model compression techniques (Choudhary *et al.*, 2020; Cai *et al.*, 2026) provide another means to reduce the model's computational footprint. Nonetheless, the quest for computationally efficient algorithms while maintaining high accuracy is an important area for future research, particularly in order to facilitate the widespread deployment of ViTs in agricultural settings with limited computational resources.

Large Amount of Model Parameters in Plant and Animal Images: As compared to CNNs, ViTs have a large number of parameters, which makes them more complicated. The more complicated the model, the more computational resources are needed. For example, to apply ViTs in real-time weed detection in the field and real-time continuous livestock health monitoring, we need light models that can run on edge devices without consuming too many computational resources (Haurum *et al.*, 2023; Wei *et al.*, 2023). However, reducing the number of parameters leads to a performance drop. Therefore, it is worthwhile to explore lightweight ViT architectures with less parameters but high accuracy.

Environmental Variability in the Field and Farm: Agricultural and veterinary detection environments are varying, which is caused by the influence of weather, lighting, soil, camera angle, etc. These factors will cause the instable detection result. For example, different lighting condition will influence the weed detection model in the field (Chen *et al.*, 2024). Therefore, it is important to make the model more robust in the face of the environmental changes. Different data augmentation and domain adaptation methods can improve the robustness of the model (Xiao *et al.*, 2024; Gao *et al.*, 2024c; Liu *et al.*, 2025; Wang *et al.,* 2025). However,

more researches are needed to develop more effective methods to make the model be adapted to the dynamic environment.

Small Object Detection in Plant and Animal Imaging: In many application scenarios, it is necessary to detect small objects. For example, in plant diseases, it is necessary to detect the initial symptoms of crop diseases; in veterinary medicine, it is necessary to detect the health status of livestock. These small objects occupy a few pixels in high resolution images. It is difficult to detect these small objects in high resolution images. Although ViTs can capture long range dependencies and improve the small object detection, ViTs still have difficulty in extracting detailed features, such as edges, and multi-scale information (Rekavandi *et al.*, 2023). Therefore, it is necessary to explore more effective methods to enhance the ability of ViTs to extract fine-grained information, and improve the performance of small object detection by designing channel-wise self-attention mechanism and multi-scale feature fusion (Dubey *et al.*, 2022; Xu *et al.*, 2023).

Data Annotation and Labeling in Plant and Animal Sciences: Access to high-quality annotated data remains a critical challenge in plant and animal sciences, because objects and backgrounds are complicated and diverse. For example, labeling different stages of plant diseases or classifying similar appearances of weeds require domain knowledge and large number of efforts. To alleviate such reliance on large labeled data, designing efficient data annotation methods and learning from semi-supervised or self-supervised methods are promising solutions (Zhao *et al.*, 2024). Furthermore, transfer learning and few-shot learning methods can also be explored to improve model performance with limited labeled data (Iman *et al.*, 2023; Lin *et al.*, 2025).

Conclusions: Our analysis of 30 studies confirms the transformative potential of ViTs in animal and plant object detection. ViTs have proven exceptionally capable in complex reasoning and feature extraction, enabling advances in tasks from crop monitoring to livestock surveillance. However, their practical deployment is currently constrained by challenges such as high computational demands, limited robustness in variable environments, and difficulties with small objects. Moving forward, the field must prioritize the development of more efficient, lightweight models and invest in the creation of comprehensive, high-quality datasets. Success in these areas, supported by closer collaboration between computer science and agricultural sciences, is crucial for integrating ViTs effectively into the future of digital farming.

Data Availability No datasets were generated or analysed during the current study.

Competing Interests The authors declare no competing interests.

Declaration Of Generative Ai in Preparation of Manuscript During the preparation of this work, the authors used “Deepseek” to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the publication's content.

Fundings This work is supported by the Natural Science Foundation of Fujian Province of China No.2022J01821, 2022J05163 and 2022J01806. The National Natural Science Foundation of China No. 11705068 and 32202219. The Foundation of Fujian Educational Committee of China No. JAT220188, Fujian University Alliance of Physics Discipline No. FJPHYS-2022-B09, Undergraduate Education Reform Project of Jimei University No. JG21082 and Teaching Reform Project of Ideological Education of Jimei University No. KCSZ077

REFERENCES

Apostolopoulos, I. D., M. Tzani and S. I. Aznaouridis (2023). A general machine learning model for assessing fruit quality using deep image features. AI 4: 812-830. doi:https://doi.org/10.3390/ai4040041

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko. (2020). *End-to-end object detection with transformers.* Paper presented at the European Conference on Computer Vision, 213-229, Virtual. doi:https://doi.org/10.48550/arXiv.2005.12872

Chen, X., C. Wu, P. Dang, Y. Chen, C. Tang and L. Qi (2024). Recognizing weed in rice field using ViT-improved YOLOv7. Transactions of the CSAE 40: 185-193. doi:http://tcsae.org/en/article/doi/10.11975/j.issn.1002-6819.202310004

Choudhary, T., V. Mishra, A. Goswami and J. Sarangapani (2020). A comprehensive survey on model compression and acceleration. Artif. Intell. Rev. 53: 5113-5155. doi:https://doi.org/10.1007/s10462-020-09816-7

Chu, X., Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia and C. Shen. (2021). *Twins: Revisiting the design of spatial attention in vision transformers.* Paper presented at the 35th Conference on Neural Information Processing Systems, 9355-9366, Virtual. doi:https://doi.org/10.48550/arXiv.2104.13840

Cai, Y., M. Lin, Z. Gao, H. Cai and H. Ni (2026). Edge device–oriented tomato fruit thinning and harvesting model under adverse weather conditions. Physiol. Plantarum 178: e70764.

doi:https://doi.org/10.1111/ppl.70764

da Silva Ferreira, M. V., S. Barbon Junior, V. G. Turrisi da Costa, D. F. Barbin and J. Lucena Barbosa Jr (2024). Deep computer vision system and explainable artificial intelligence applied for classification of dragon fruit (Hylocereus spp.). Sci. Hortic. 338: 113605. doi:https://doi.org/10.1016/j.scienta.2024.113605

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold and S. Gelly (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. doi:https://doi.org/10.48550/arXiv.2010.11929

Dubey, S., F. Olimov, M. A. Rafique and M. Jeon (2022). Improving small objects detection using transformer. J. Visual Commun. Image Represent. 89: 103620. doi:https://doi.org/10.1016/j.jvcir.2022.103620

Dümen, S., E. Kavalcı Yılmaz, K. Adem and E. Avaroglu (2024). Performance of vision transformer and swin transformer models for lemon quality classification in fruit juice factories. Eur. Food Res. Technol. 250: 2291-2302. doi:https://doi.org/10.1007/s00217-024-04537-5

Fu, H., X. Zhao, H. Wu, S. Zheng, K. Zheng and C. Zhai (2022). Design and experimental verification of the yolov5 model implanted with a transformer module for target-oriented spraying in cabbage farming. Agronomy 12: 2551. doi:https://doi.org/10.3390/agronomy12102551

Fu, Z., L. Yin, C. Cui and Y. Wang (2024). A lightweight MHDI-DETR model for detecting grape leaf diseases. Front. Plant Sci. 15: 1499911. doi:https://doi.org/10.3389/fpls.2024.1499911

Gao, Z., J. Huang, J. Chen, T. Shao, H. Ni and H. Cai (2024a). Deep transfer learning-based computer vision for real-time harvest period classification and impurity detection of Porphyra haitnensis. Aquacult. Int. 32: 5171-5198. doi:https://doi.org/10.1007/s10499-024-01422-6

Gao, Z., S. Chen, J. Huang and H. Cai (2024b). Real‐time quantitative detection of hydrocolloid adulteration in meat based on Swin Transformer and smartphone. J. Food Sci. 89: 4359-4371. doi:https://doi.org/10.1111/1750-3841.17159

Gao, Z., Q. Lin, Q. He, C. Liu, H. Cai and H. Ni (2024c). Rapid detection of spoiled apple juice using electrical impedance spectroscopy and data augmentation-based machine learning. Chiang Mai J. Science 51: e2024071. doi:https://doi.org/10.12982/CMJS.2024.071

Geng, H., Z. Hou, J. Liang, X. Li, X. Zhou and A. Xu (2024). Motion focus global‐local network: Combining attention mechanism with micro action features for cow behavior recognition. Comput. Electron. Agric. 226: 109399. doi:https://doi.org/10.1016/j.compag.2024.109399

Graham, B., A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou and M. Douze. (2021). *Levit: a vision transformer in convnet's clothing for faster inference.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 12259-12269, Virtual. doi: https://doi.org/10.48550/arXiv.2104.01136

Guo, Y., W. Hong, J. Wu, X. Huang, Y. Qiao and H. Kong (2023). Vision-based cow tracking and feeding monitoring for autonomous livestock farming: the YOLOv5s-CA+ DeepSORT-vision transformer. IEEE Robot. Autom. Mag. 30: 68-76. doi:https://doi.org/10.1109/MRA.2023.3310857

Han, K., Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu and Y. Xu (2022). A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45: 87-110. doi:https://doi.org/10.1109/TPAMI.2022.3152247

Hatamizadeh, A., G. Heinrich, H. Yin, A. Tao, J. M. Alvarez, J. Kautz and P. Molchanov (2023). Fastervit: Fast vision transformers with hierarchical attention. arXiv preprint arXiv:2306.06189. doi:https://doi.org/10.48550/arXiv.2306.06189

Haurum, J. B., S. Escalera, G. W. Taylor and T. B. Moeslund. (2023). *Which tokens to use? investigating token reduction in vision transformers.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 773-783, Paris, France. doi:https://doi.org/10.48550/arXiv.2308.04657

He, Y., N. Zhang, X. Ge, S. Li, L. Yang, M. Kong, Y. Guo and C. Lv (2025). Passion fruit disease detection using sparse parallel attention mechanism and optical sensing. Agriculture 15: 733. doi:https://doi.org/10.3390/agriculture15070733

Iman, M., H. R. Arabnia and K. Rasheed (2023). A review of deep transfer learning and recent advancements. Technologies 11: 40. doi:https://doi.org/10.3390/technologies11020040

Jamali, A., B. Lu, E. M. Gerbrandt, C. Teasdale, R. R. Burlakoti, S. Sabaratnam, J. Mcintyre, L. Yang, M. Schmidt and D. Mccaffrey (2025). High-resolution UAV-based blueberry scorch virus mapping utilizing a deep vision transformer algorithm. Comput. Electron. Agric. 229: 109726.

doi:https://doi.org/10.1016/j.compag.2024.109726

Koike, M., R. Onuma, R. Adachi and H. Mineno (2024). Transformer-based water stress estimation using leaf wilting computed from leaf images and unsupervised domain adaptation for tomato crops. Technologies 12: 94. doi:https://doi.org/10.3390/technologies12070094

Kumar, P., S. Luo and K. Shaukat (2023). A comprehensive review of deep learning approaches for animal detection on video data. IJACSA 14: 11. doi:https://doi.org/10.14569/IJACSA.2023.0141144

Lamping, C., G. Kootstra and M. Derks (2025). Transformer-based similarity learning for re-identification of chickens. Smart Agric. Technol. 11: 100945. doi:https://doi.org/10.1016/j.atech.2025.100945

Li, J., B. Yang, J. Chen, J. Liu, F. K. Amevor, G. Chen, B. Zhang and X. Zhao (2025). PoulTrans: a transformer-based model for accurate poultry condition assessment. Sci. Rep. 15: 14064. doi:https://doi.org/10.1038/s41598-025-98078-w

Lin, Y., Y. Cai, H. Chen, Y. Cai, Z. Lin, H. Cai and H. Ni (2025). Adaptive feedback cross-loop for preserving and robust spectral information optimization without spectral processing in few-shot learning. Meas. Sci. Technol. 36: 075503. doi:https://doi.org/10.1088/1361-6501/aded2a

Liu, W., Y. Lin, Y. Cai, H. Cai and H. Ni (2025a). Detecting Adulterants in Tea Using Mid-Infrared Spectroscopy: A Comparative Study of Deep Learning and Machine Learning. J. Anim. Plant Sci. 35: 889-899. doi:https://doi.org/10.36899/JAPS.2025.4.0077

Liu, X., Q. Sui and Z. Chen (2025). Real time weed identification with enhanced mobilevit model for mobile devices. Sci. Rep. 15: 27323. doi:https://doi.org/10.1038/s41598-025-12036-0

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo. (2021). *Swin transformer: Hierarchical vision transformer using shifted windows.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022, Virtual. doi:https://doi.org/10.48550/arXiv.2103.14030

Lu, J., L. Tan and H. Jiang (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. Agriculture 11: 707. doi:https://doi.org/10.3390/agriculture11080707

Mehta, S. and M. Rastegari (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178. doi:https://doi.org/10.48550/arXiv.2110.02178

Ni, X., F. Wang, H. Huang, L. Wang, C. Wen and D. Chen (2024). A cnn-and self-attention-based maize growth stage recognition method and platform from uav orthophoto images. Remote Sens. 16: 2672. doi:https://doi.org/10.3390/rs16142672

Pei, G., X. Qian, B. Zhou, Z. Liu and W. Wu (2025). Research on agricultural disease recognition methods based on very large Kernel convolutional network-RepLKNet. Sci. Rep. 15: 16843. doi:https://doi.org/10.1038/s41598-025-01553-7

Rekavandi, A. M., S. Rashidi, F. Boussaid, S. Hoefs and E. Akbas (2023). Transformers in small object detection: A benchmark and survey of state-of-the-art. arXiv preprint arXiv:2309.04902. doi:https://doi.org/10.1145/375809

Shimazu, R., C. S. Leow, P. Buayai, X. Mao, W.-Y. Chung and H. Nishizaki (2025). Non-invasive estimation of Shine Muscat grape color and sensory evaluation from standard camera images. The Vis. Comput. doi:https://doi.org/10.1007/s00371-025-03925-6

Shuai, Y., J. Shi, Y. Li, S. Zhou, L. Zhang and J. Mu (2025). YOLO-SW: A real-time weed detection model for soybean fields using Swin Transformer and RT-DETR. Agronomy 15: 1712. doi:https://doi.org/10.3390/agronomy15071712

Singh, A. K., A. Rao, P. Chattopadhyay, R. Maurya and L. Singh (2024). Effective plant disease diagnosis using Vision Transformer trained with leafy-generative adversarial network-generated images. Expert Syst. Appl. 254: 124387. doi:https://doi.org/10.1016/j.eswa.2024.124387

Sui, J., L. Liu, Z. Wang and L. Yang (2025). RE-YOLO: An apple picking detection algorithm fusing receptive-field attention convolution and efficient multi-scale attention. PLoS One 20: e0319041. doi:https://doi.org/10.1371/journal.pone.0319041

Vaswani, A. (2017). *Attention is all you need.* Paper presented at the the 31st International Conference on Neural Information Processing Systems, 6000-6010, Long Beach, California, USA. doi:https://doi.org/10.48550/arXiv.1706.03762

Wang, W., E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo and L. Shao. (2021). *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vsion,

568-578, Virtual. doi: https://doi.org/10.1109/ICCV.2021.00061

Wang, Y., H. Ke and H. Cai (2025). PC-YOLO: Enhancing Object Detection in Adverse Weather. J. Electron. Imaging 34: 023049. doi:https://doi.org/10.1117/1.JEI.34.2.023049

Wei, S., T. Ye, S. Zhang, Y. Tang and J. Liang. (2023). *Joint token pruning and squeezing towards more aggressive compression of vision transformers.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2092-2101, Vancouver, Canada. doi: https://doi.org/10.1109/CVPR52729.2023.00213

Xiao, Y., H. Cai and H. Ni (2024). Identification of geographical origin and adulteration of Northeast China soybeans by mid-infrared spectroscopy and spectra augmentation. J. Consum. Prot, Food Saf. 19: 99-111. doi:https://doi.org/10.1007/s00003-023-01471-8

Xu, S., J. Gu, Y. Hua and Y. Liu (2023). Dktnet: dual-key transformer network for small object detection. Neurocomputing 525: 29-41. doi:https://doi.org/10.1016/j.neucom.2023.01.055

Xu, Z., Y. Zhao, Z. Yin and Q. Yu (2024). Optimized BottleNet Transformer model with Graph Sampling and Counterfactual Attention for cow individual identification. Comput. Electron. Agric. 218: 108703. doi:https://doi.org/10.1016/j.compag.2024.108703

Zhang, J., D. Li, X. Shi, F. Wang, L. Li and Y. Chen (2025). DCTnet: a double-channel transformer network for peach disease detection using UAVs. Complex Intell. Syst. 11: 111. doi:https://doi.org/10.1007/s40747-024-01749-w

Zhang, S., H. Liu, S. Lin and K. He. (2024). *You only need less attention at each stage in vision transformers.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6057-6066, Seattle, USA. doi:https://doi.org/10.1109/CVPR.2024.00213

Zhao, J., P. Zeng, G. Shen, Q. Chen and M. Guo (2024). Hardware‐software co-design enabling static and dynamic sparse attention mechanisms. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 43: 2783-2796. doi:https://doi.org/10.1109/TCAD.2024.3373592

Zhao, S., Z. Bai, S. Wang and Y. Gu (2023). Research on automatic classification and detection of mutton multi-parts based on swin-transformer. Foods 12: 1642. doi: https://doi.org/10.3390/foods12081642

Zhao, Z., L. Alzubaidi, J. Zhang, Y. Duan and Y. Gu (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Syst. Appl. 242: 122807. doi:https://doi.org/10.1016/j.eswa.2023.122807

Zhu, X., W. Su, L. Lu, B. Li, X. Wang and J. Dai (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159. doi:https://doi.org/10.48550/arXiv.2010.04159.