

# A SYSTEMATIC REVIEW OF TRANSFORMER-BASED VISION MODELS FOR OBJECT DETECTION IN FOOD AND AGRICULTURE

Maolan Lin<sup>1</sup>, Zhenchang Gao<sup>2</sup>, Wenliang Liao<sup>1</sup>, Honghao Cai<sup>1</sup>✉

<sup>1</sup>Department of Physics, School of Science,  
Jimei University, **China**

<sup>2</sup>School of Information Science and Technology, ShanghaiTech University,  
Shanghai, **China**

## ABSTRACT

Computer vision has become a cornerstone technology in the food and agriculture industries, driving innovation and enabling automation across a wide range of processes. Within this field, object detection plays a critical role, supporting efficiency, accuracy, and scalability in real-world applications. The Transformer, first introduced in natural language processing, demonstrated outstanding performance thanks to its powerful self-attention mechanism and parallel processing capabilities. More recently, it has been rapidly adopted in object detection and is emerging as a strong alternative to traditional convolutional neural networks. However, much of the related research remains scattered and interdisciplinary. This paper systematically reviews the development of transformer-based models for computer vision, analysing research trends, key topics, and distinctions from other algorithms. It introduces the basic architecture of the Vision Transformer (ViT) and other transformer-based vision models, explains core principles such as self-attention and multi-stage processing, and examines applications in food and agriculture, including food quality analysis, crop monitoring, pest and disease detection, and weed identification. Challenges and future directions of transformer-based models are also discussed, alongside a review of the latest research for reference. By consolidating a large body of literature, this study provides a comprehensive overview of the structure, development, advantages, and limitations of transformer-based vision models, while highlighting their potential to deliver more intelligent, sustainable, and efficient decision-support systems for precision food and farming practices.

**Keywords:** deep learning, convolutional neural network, computer vision, intelligent agriculture, self-attention mechanism, food quality

## INTRODUCTION

The rapid development of digital technologies has transformed the food and agriculture sector, with computer vision playing a central role in improving efficiency and productivity (Patrício and Rieder, 2018; Singh et al., 2021). Applications include chemical and physical property detection (Xiao et al., 2024b; Harnsoongnoen and Jaroensuk, 2021), quality classification (Wang, 2022), adulteration and counterfeit detection (Gao et

al., 2024c), pest and disease recognition (Abbaspour-Gilandeh et al., 2022), precise fertilisation and spraying (Ghazal et al., 2024), optimised irrigation, accurate harvesting, and real-time crop monitoring (Li et al., 2021b; Sharma et al., 2021). These technologies have enhanced efficiency, improved quality, reduced labour costs, and boosted yields. At the core of many of these innovations lies object detection, which simultaneously

✉hhcai@jmu.edu.cn, <https://orcid.org/0000-0002-1870-8061>

classifies and localises objects within an image (Zhao et al., 2019). Falling hardware costs and greater computational power have fuelled rapid adoption across domains such as autonomous driving (Bochkovskiy, 2020), medical imaging (Zhou et al., 2021), video surveillance (Aradhya, 2019), unmanned aerial vehicles (Li et al., 2019), food quality detection (Bianco et al., 2023) and agricultural (Zhang et al., 2020).

A range of neural network architectures has advanced visual recognition. For instance, the Faster Region-Based Convolutional Neural Network (Faster R-CNN) achieves fine-grained object classification and bounding box regression through region proposal networks (Ren et al., 2017). The Single Shot Multibox Detector (SSD) predicts both the locations and classes of multiple objects across different scales in an image, allowing efficient object detection within a single network (Ning et al., 2017). RetinaNet addresses class imbalance using focal loss (Lin et al., 2017), while the Mask Region-Based Convolutional Neural Network (Mask R-CNN) extends Faster R-CNN by enabling instance segmentation with pixel-level masks, giving it an advantage in tasks that demand fine-grained object segmentation (He et al., 2017). Among these, the You Only Look Once (YOLO) algorithm stands out for its real-time performance, treating object detection as a regression problem by predicting bounding boxes and class labels simultaneously in a single neural network. Successive versions of YOLO have progressively advanced its capabilities. YOLOv2 and YOLOv3 introduced multiscale detection and enhanced feature extraction, while YOLOv4 further improved accuracy with stronger multi-scale detection capability. YOLOv5 emphasised model simplification and optimisation, resulting in a lightweight and fast architecture well-suited for real-time applications requiring efficient inference and cross-platform deployment (Jiang et al., 2022b; Redmon et al., 2016). More recently, YOLO has evolved to YOLOv11, which strengthens feature extraction and multitask processing, achieving state-of-the-art performance across diverse computer vision tasks. It offers multiple model sizes, ranging from nano to extra-large, enabling a balance between precision and computational efficiency, making it adaptable to both resource-constrained edge devices and high-performance computing environments (Khanam and Hussain, 2024).

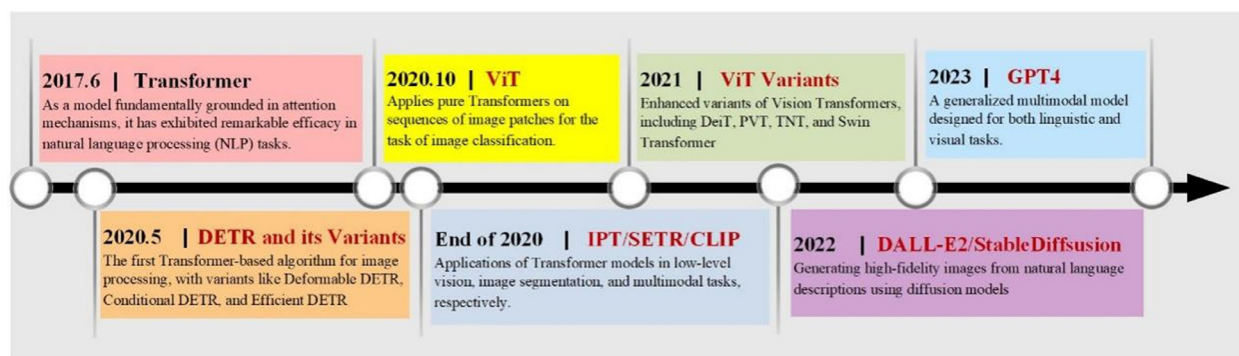
However, as data volumes grow and models become increasingly complex, the limitations of traditional algorithms have become more evident (Li et al., 2024b; Zhao et al., 2019). For example, convolutional neural networks (CNNs) rely primarily on local convolution operations, which constrain their ability to capture long-range contextual dependencies within an image (Krizhevsky et al., 2017). Recurrent neural networks (RNNs) also struggle to effectively model spatial information, particularly when applied to large-scale images (Lipton et al., 2015). Similarly, the YOLO series (e.g., YOLOv1 and YOLOv2) face challenges in detecting small objects and achieving precise localisation (Redmon et al., 2016; Redmon, 2018).

Transformers, first introduced in 2017 for natural language processing (NLP) (Vaswani et al., 2017), have since revolutionised computer vision through their powerful self-attention mechanism, which captures both global context and long-range dependencies. Fig. 1 presents the development timeline of transformer-based vision models. In May 2020, Detection Transformer (DETR) was the first to apply transformers to object detection tasks, significantly improving accuracy while eliminating many manually designed components prevalent in earlier methods, such as anchor boxes and non-maximum suppression (Carion et al., 2020). Since then, numerous DETR variants have emerged. For example, Deformable DETR enhanced the detection of small and dense objects by incorporating deformable convolutions into the architecture (Zhu et al., 2020a). Conditional DETR proposed a conditional cross-attention mechanism to mitigate slow training convergence (Meng et al., 2021). Efficient DETR optimised the architecture and training strategies to improve efficiency in resource-constrained environments (Yao et al., 2021). In October 2020, the Vision Transformer (ViT) became the first pure transformer architecture designed specifically for image classification (Dosovitskiy et al., 2021). ViT splits images into fixed-size patches, treats each patch as an input token, and learns global features through self-attention. This approach demonstrated exceptional performance on large datasets, such as ImageNet, confirming the effectiveness of transformers for image classification. By the end of 2020, the Image Processing Transformer (IPT) emerged, offering improved efficiency and accuracy in image

processing tasks, particularly with complex and high-resolution data (Chen et al., 2021). Around the same time, the Semantic Segmentation Transformer (SETR) advanced semantic segmentation accuracy (Strudel et al., 2021; Xie et al., 2021). Moreover, the Contrastive Language-Image Pretraining (CLIP) model enabled zero-shot transfer learning across diverse vision-language tasks by jointly learning images and text representations (Li et al., 2021a). In 2021, multiple ViT variants were introduced. The Data-efficient Image Transformer (DeiT) employed knowledge distillation to reduce the reliance on large-scale datasets, achieving strong performance on smaller datasets (Touvron et al., 2021). Pyramid ViT used a pyramid structure to better handle multiscale image information by constructing feature maps at multiple levels (Wang et al., 2021b). The Transformer-in-Transformer (TNT) architecture nested transformers within one another at different layers, enhancing feature learning for image classification and related tasks (Han et al., 2021). The Swin Transformer addressed the high computational cost of large-scale image processing by introducing a shifted-window mechanism (Liu et al., 2021). In 2022, two notable generative models emerged: DALL-E 2, a high-quality image generation model developed by OpenAI, and Stable Diffusion, an efficient text-to-image model designed for optimised resource usage (Cao et al., 2024; Zhang et al., 2023). In 2023, OpenAI released GPT-4, a large language model built on the transformer architecture (Achiam et al., 2023; Vaswani et al., 2017). Through large-scale pretraining and generative tasks, GPT-4 has achieved significant advancements in understanding and generating text,

greatly enhancing productivity and innovation across diverse fields.

Object detection in food and agriculture spans diverse spatial and temporal scales and often occurs in unstructured, complex environments (Marinoudi et al., 2019). A variety of deep learning models have been employed in this field: CNNs for food defect detection (Zhu et al., 2021), deep CNNs for fruits and vegetable freshness assessment (Fahad et al., 2022), Faster R-CNN for crop classification and disease detection (Gong and Zhang, 2023; Mu et al., 2022), SSD for weed-crop separation (Cai et al., 2020; Zhao et al., 2021), and Mask R-CNN for disease segmentation (Jabir et al., 2023). U-Net has been used for crop health monitoring and soil management (Wang et al., 2021a), while deep CNNs have addressed pest and disease detection (Ale et al., 2019). Deep RNNs have supported fruit classification (Ndikumana et al., 2018), and YOLO has been applied to small object detection, crop growth evaluation, and real-time pest and disease identification (Mostafa et al., 2024; Rajamohanam and Latha, 2023). Progress in this area is constrained by the shortage of labelled, publicly available datasets. The absence of high-quality annotations limits training effectiveness, hinders performance improvements, and ultimately diminishes detection accuracy. Practical deployment presents further challenges, including adaptation to diverse environments, real-time processing requirements, and robust recognition in cluttered backgrounds (Ariza-Sentís et al., 2024). Transformers address many of these limitations. They effectively manage long-range dependencies, process large datasets, and minimise inductive bias. Their self-attention



**Fig. 1.** Key milestones in the development of transformers

mechanism enables parallelisation across sequences, in contrast to recurrent architectures, such as RNNs or Long Short-Term Memory Networks (LSTMs), that rely on sequential processing. This parallelisation substantially accelerates training, especially for large datasets or long sequences (Vaswani et al., 2017), allowing transformers to achieve state-of-the-art efficiency and accuracy. Advances in efficient transformer architectures have also led to increasingly lightweight models (Touvron et al., 2021). Moreover, their ability to integrate multimodal data – for example, combining optical and infrared imagery – offers significant advantages in multisource agricultural applications (Khan et al., 2022; Shipitko et al., 2020).

In recent years, transformer-based object detection in food and agriculture has advanced rapidly, though research outputs remain fragmented and uneven in rigour. This paper seeks to consolidate these developments by presenting a comprehensive review of transformer-based vision applications in the field. We analyse the strengths and limitations of transformer-based detectors, examine their integration with

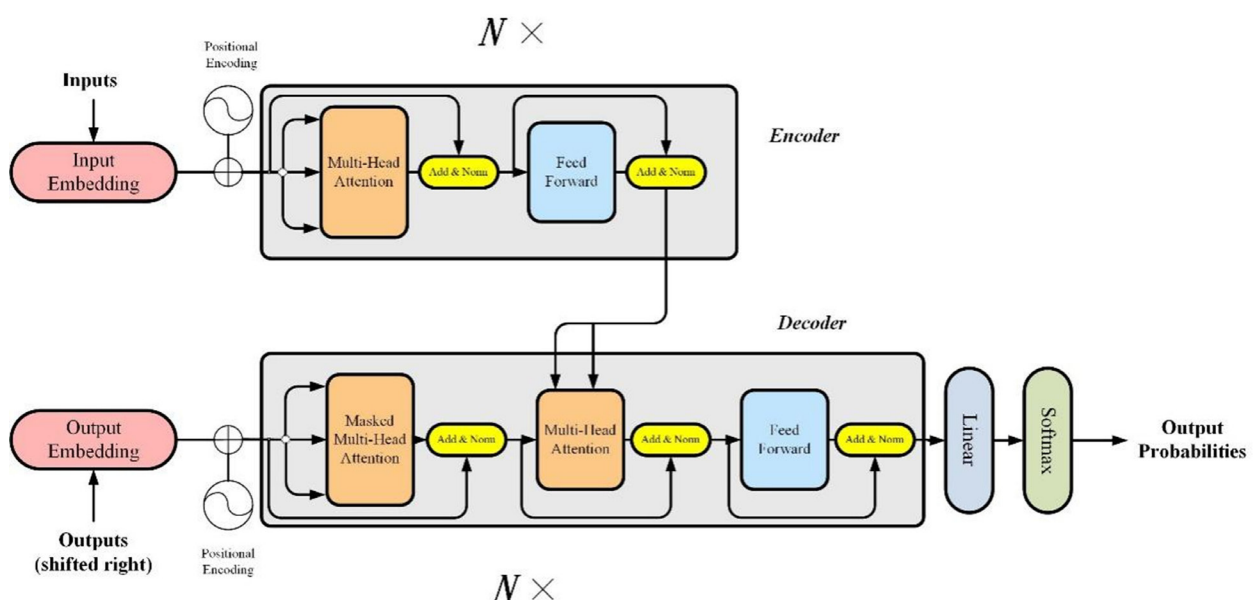
advanced algorithms, and assess their potential and challenges for agricultural object detection.

The remainder of this paper is organised as follows: Section 2 introduces the transformer architecture; Section 3 reviews applications of transformer-based models in food and agricultural object detection; Section 4 discusses key challenges; and Section 5 outlines future research directions.

## TRANSFORMER ARCHITECTURE

### Background and motivation

The Transformer architecture introduced by Vaswani et al. (2017) revolutionised sequence modelling by overcoming the limitations of RNNs and LSTMs. Unlike its predecessors, the Transformer relies exclusively on self-attention mechanisms to capture dependencies within the input data, enabling efficient parallel processing and eliminating the need for recurrence. This design greatly improves training efficiency and model performance across a wide range of NLP tasks (Lee and Toutanova, 2019). As illustrated in Fig. 2,



The encoder (top) consists of stacked layers with two main sublayers: self-attention, which captures relationships within the input sequence, and a feed-forward neural network that processes these representations. The decoder (bottom) also consists of stacked layers but includes three sublayers: self-attention for the current context, encoder–decoder attention that combines encoder output with decoding context, and a feed-forward network. This design effectively captures long-range dependencies and enables parallel computation.

**Fig. 2.** Architecture of the transformer model

the Transformer model consists of three core components: Multi-Head Attention (MHA), positional encoding, and feed-forward networks. Of these, MHA and feed-forward networks form the primary building blocks of both the encoder and the decoder. Because the Transformer lacks recurrent structures, it requires a mechanism to represent the order of tokens within a sequence. Positional encoding fulfils this role by embedding positional information into each input element, typically through sine and cosine functions.

### Basic architecture

The Transformer model comprises two main components: the encoder and the decoder. As shown in Fig. 2, the encoder is built from a stack of identical encoder layers, each containing two primary sub-layers: MHA and a Feed-Forward Neural Network (FFN). Similarly, the decoder consists of a stack of identical decoder layers, each with three sub-layers: MHA, encoder–decoder attention, and FFN. Both encoder and decoder sub-layers are followed by residual connections and layer normalisation. The output of each sub-layer can be expressed as (1):

$$Out\ put = LayerNorm(x + SubLayer(x)) \quad (1)$$

where:  $x$  – is the input sequence, and  $SubLayer$  – refers to either the attention module or feedforward network.

### Self-attention mechanism

#### Scaled dot-product attention

In the self-attention mechanism, each element in a sequence can attend to every other element, allowing the model to capture relationships between words regardless of their positional distance. This process involves the following steps:

(1) Linear Transformations: Each input vector  $x_i$  is linearly transformed into three vectors: a query  $Q$ , a key  $K$ , and a value  $V$ . The input sequences are represented as  $x_i \in \mathbb{R}^{n_i \times d_{x_i}}$ , where  $n$  denotes the number of elements in the input sequence and  $d$  represents the dimensionality of each element’s vector representation. These transformations are defined as:

$$Q_i = x_i W^Q \quad (2)$$

$$K_i = x_i W^K \quad (3)$$

$$V_i = x_i W^V \quad (4)$$

Here:  $W^Q$ ,  $W^K$ , and  $W^V$  – are learned weight matrices corresponding to queries, keys, and values, respectively.

(2) Attention score layer: As shown in Fig. 3, the attention scores are computed by taking the dot product of the query vector with all key vectors, applying

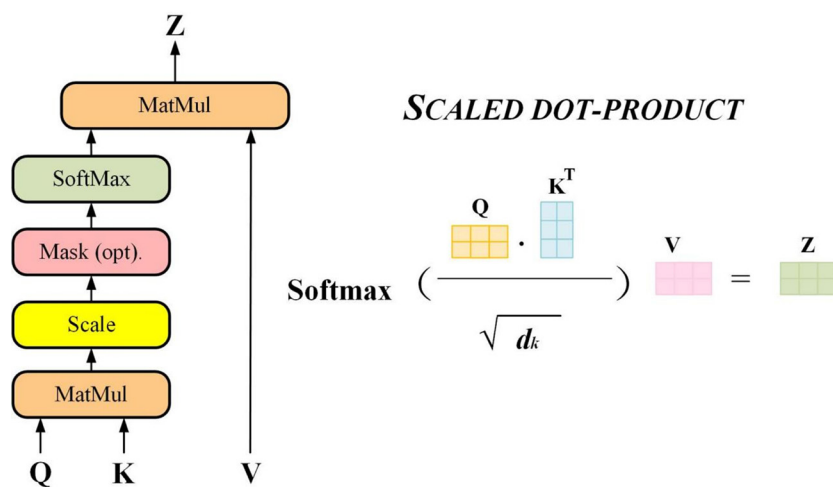


Fig. 3. The scaled dot-product

a scaling factor, and then using a *softmax* function to obtain normalised weights:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Here:  $\frac{QK^T}{\sqrt{d_k}}$  denotes the scaled dot-product attention, where  $d_k$  – is the dimension of the key vectors, *softmax* – function ensures that the resulting attention weights sum to one.

In practice, efficiency is often enhanced by processing from a mini-batch perspective. For instance, attention can be computer-based on  $n$  queries and  $m$  key–value pairs, where the query and key vectors each have length  $d$ , and the value vectors have length  $v$ . In this case, the scaled dot-product attention, given the query  $Q \in R^{n \times d}$ , key  $K \in R^{m \times d}$ , and value  $V \in R^{m \times v}$ , is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \in R^{n \times v} \quad (6)$$

(3) Weighted Sum: The resulting attention scores are then used to compute a weighted sum of the value vectors:

$$Output_i = \sum_j softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V_j \quad (7)$$

This output vector represents the context of each element  $x_i$  – by aggregating information from all other elements in the sequence, weighted by their relevance.

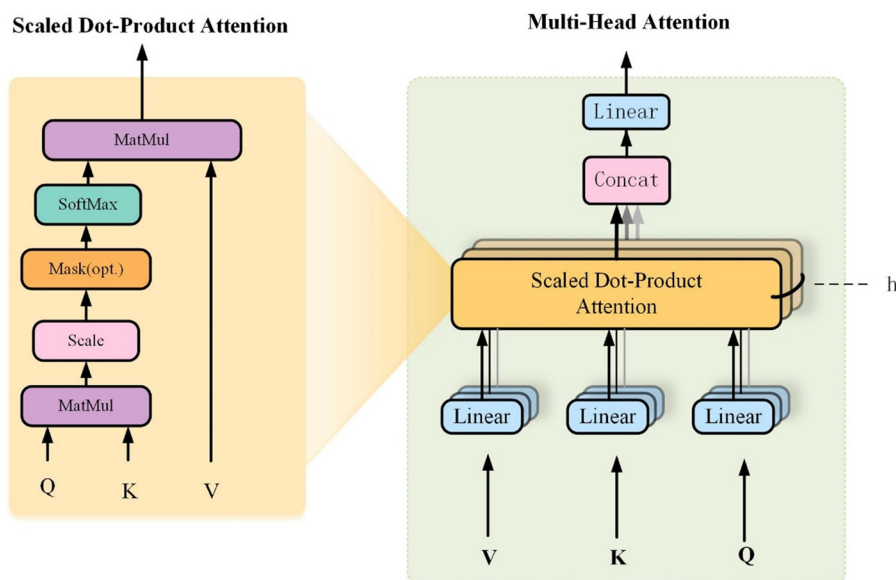
### Multi-head attention

To strengthen the model’s ability to capture diverse relationships in the data, the Transformer employs an MHA mechanism. As shown in Fig. 4 (right), this involves running multiple self-attention operations in parallel, each with its own learned linear transformations. This design increases the model’s expressive power, reduces information loss, and stabilises training. The outputs of all attention heads are concatenated and passed through a linear transformation, as shown in Equation (8):

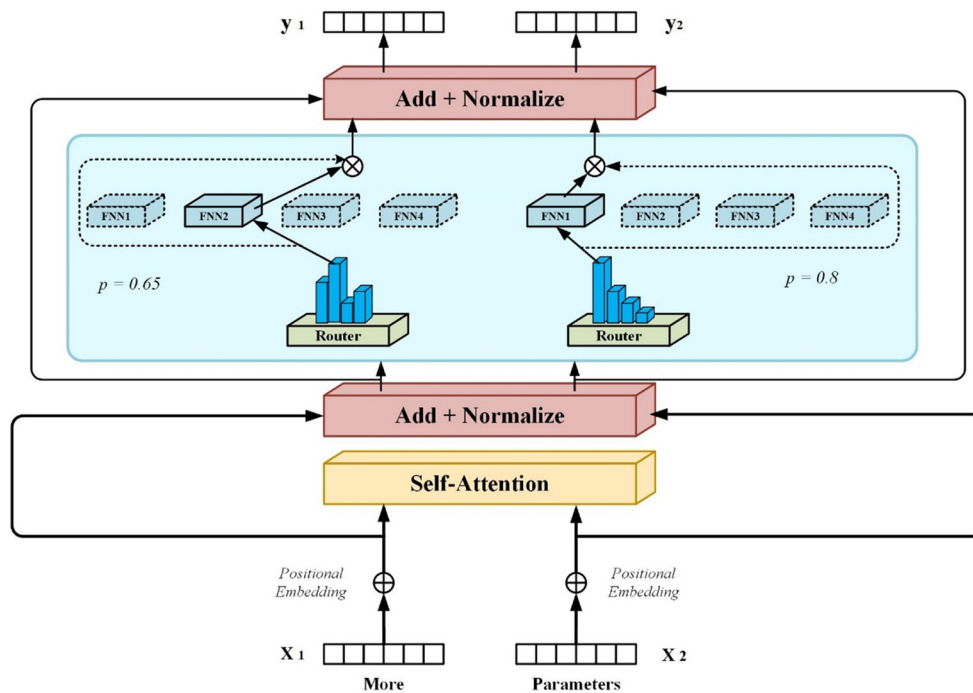
$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^o \quad (8)$$

Here, the attention mechanism is divided into  $h$  independent heads, each with its own learned weight matrices  $W^Q$ ,  $W^K$ , and  $W^V$ . The outputs of all the heads are then concatenated, and each head’s computation is defined in Equation (9):

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (9)$$



**Fig. 4.** Scaled dot-product attention (left); multi-head attention, which consists of several attention layers running in parallel (right)



**Fig. 5.** Model architecture of the position-wise feed-forward network

### Position-wise feed-forward network

The position-wise feed-forward network follows the self-attention, masked self-attention, and MHA layers. As shown in Fig. 5, it typically consists of two linear transformation layers – an input-to-hidden layer and a hidden-to-output layer. An activation function (such as ReLU) is applied between the two transformations to enhance the model’s nonlinear representation capability. Its mathematical expression is given in Equation (10):

$$FNN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (10)$$

where  $W_1$  and  $W_2$  – are weight matrices, and  $b_1$  and  $b_2$  – are bias vectors.

### Positional encoding

In the Transformer architecture, the model itself lacks an intrinsic ability to represent sequential order. To address this, an additional mechanism – positional encoding – is introduced to incorporate positional information. Its primary role is to provide each word in the input sequence with information about its position, enabling the model to capture both relative and absolute ordering.

Positional encodings are added to word embeddings to form the final representation, allowing the self-attention mechanism to account for sequence structure. By supplying this positional context, the Transformer can effectively process sequential data. Correct implementation of positional encoding is therefore essential to fully leverage the model’s capabilities.

There are two main types of positional encoding: sinusoidal and cosine. Their calculation is shown in Equation (11):

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (11)$$

where:  $pos$  – is the position index of a word in the sequence,  $i$  – denotes the dimension index, and  $d$  – is the total dimension of the encoding. This method generates smooth, unique encodings for each position, helping the model learn and utilise positional information in sequential data.

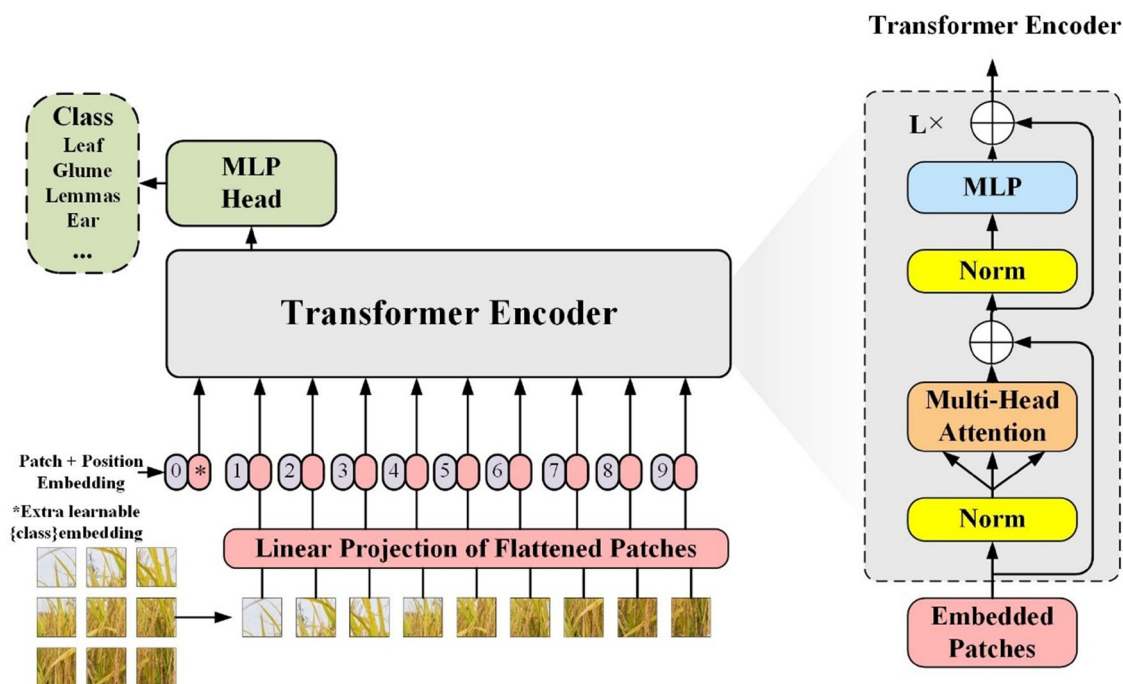


Fig. 6. Vision transformer architecture for agriculture

### Expansion and improvement

The success and widespread application of the Transformer model architecture – particularly its superior performance in NLP – has garnered significant attention. Given the limitations of CNNs in capturing long-range dependencies and modelling global context in large-scale images, the self-attention mechanism of the Transformer was first transferred to computer vision in October 2020, in the Vision Transformer (ViT) model (Dosovitskiy et al., 2021). Its global attention mechanism and parallel computation capabilities represent important advancements in the field.

### Structures newly added to ViT

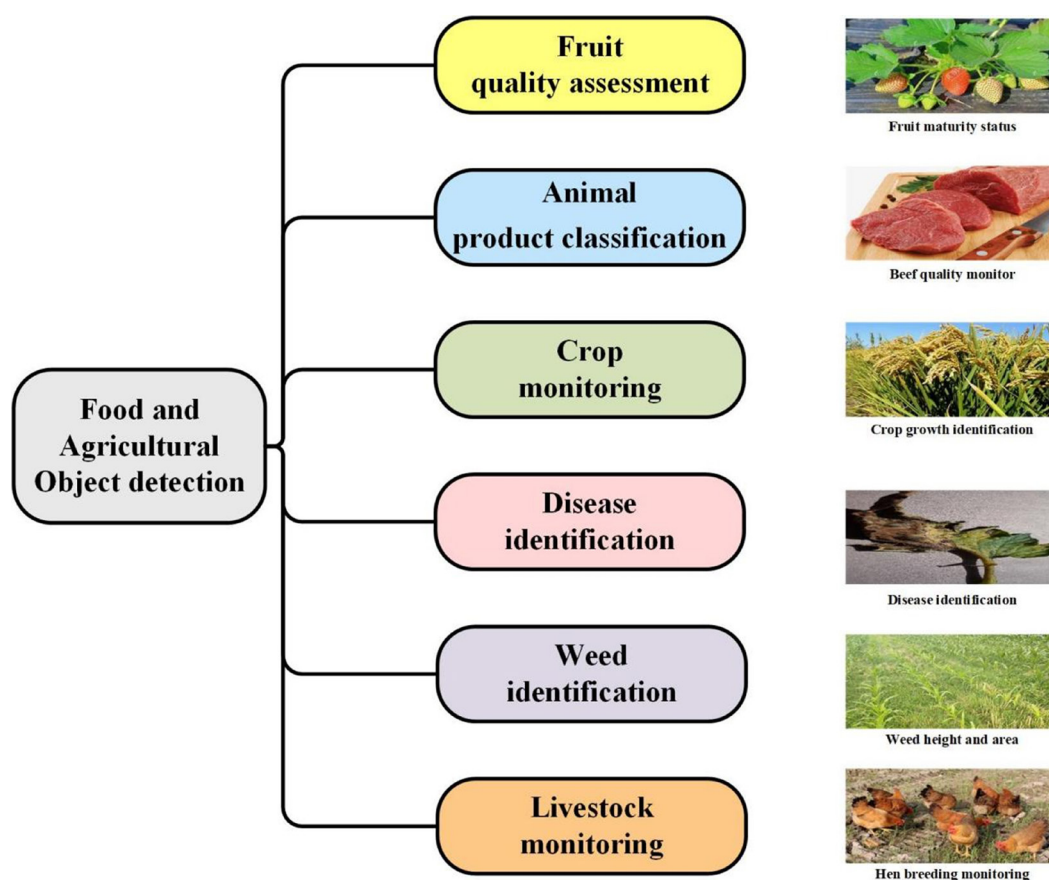
As shown in Fig. 6, ViT divides the input image into fixed-size patches, such as  $16 \times 16$  pixels. For an image of size  $H \times W$ , this produces  $N = H \times W / p^2$  patches. Each patch is then flattened and passed through a linear layer to map it into an embedding space, generating a vector of dimension  $D$ , as shown in Equation (12):

$$z_i = \text{PatchEmbedding}(x_i) = x_{ij}W + b \quad (12)$$

The patches are then processed in parallel through the Transformer’s self-attention mechanism. Finally, positional encoding is added to preserve the spatial information of the image patches, forming the final input representations.

### VISION TRANSFORMER IN FOOD AND AGRICULTURE

Transformer-based Vision models and their variants demonstrate considerable potential in food and agriculture applications, as illustrated in Fig. 7. These include fruit quality assessment, animal product classification, crop monitoring, plant disease detection, weed identification, livestock monitoring, and integrated applications. Compared with traditional manual monitoring methods and previously dominant CNN-based architectures, transformer-based models improve detection accuracy and timeliness by effectively handling long-distance dependencies and capturing global details. In this subsection, we review the current literature on the application of Vision Transformers and their variants across diverse agricultural and food-related contexts.



**Fig. 7.** Common applications of vision transformers in food and agricultural object detection

### Fruit quality assessment

Accurate fruit ripeness assessment is essential for optimising harvest timing, improving fruit quality, and reducing post-harvest losses (Khan et al., 2023). By reliably determining fruit quality, growers can maximise yields and minimise waste. Integrating advanced technologies such as Transformer-based vision models (e.g., ViTs, Swin Transformer) into fruit quality assessment enables more precise and timely evaluations of quality characteristics. This technological advancement not only supports decision-making for optimal harvest timing but also enhances sustainability in fruit production by improving food safety and reducing resource waste.

In real-world scenarios, complex lighting and weather conditions, combined with small and often obscured target objects, make assessment challenging. Although CNNs have dominated semantic segmentation in food

and agriculture, their limitations in capturing global information and managing complex scenes can reduce segmentation accuracy. To address this, researchers proposed the Powerful Decoder SETR Network (PD-SegNet) in 2023 (Zhu et al., 2023). Designed specifically for complex background scenarios, PD-SegNet balances segmentation accuracy and computational efficiency by combining dynamic kernel self-updating with edge-aware optimisation. Experimental results demonstrated that the algorithm performed exceptionally well in segmenting apple blossoms and fruits, achieving new state-of-the-art results on the apple segmentation dataset.

The use of transformer models with global receptive fields to improve efficiency and accuracy in object detection and recognition has also expanded beyond apple quality detection. For instance, a fruit dataset was constructed to train, test, and evaluate multiple

transformer models (Xiao et al., 2023). The models included ViT, Swin Transformer, and their effectiveness in assessing fruit quality was analysed. Results showed that the Swin Transformer outperformed the ViT in classifying pears and apples, enabling more accurate quality classification across multiple fruit qualities.

Quality assessment extends beyond classification to include grading within specific fruit varieties. Knott et al. (2023) proposed a machine learning approach based on pre-trained Vision Transformers for fruit classification and grading, particularly for apple defect detection and banana ripeness estimation. This method effectively distinguishes grades and supports quality classification without requiring large training datasets. Its classification accuracy was comparable to – or in some cases exceeded – the best-performing CNN models, with an error margin of no more than 1%. Moreover, the approach reduced the required number of training samples by a factor of three, making it faster and more efficient. These results highlight the effectiveness and efficiency of pre-trained Vision Transformers in food and agricultural product classification and grading.

### **Animal product classification**

Accurate detection and classification of animal product quality are critical for achieving sustainability in food and agriculture. Traditional quality assessment methods – such as chemical analysis, microbial culture, and manual physical inspection – often suffer from inefficiency and inaccuracies. Moreover, these methods generally fail to capture real-time changes in food quality and condition, which can compromise food safety in unpredictable ways. By contrast, the adoption of transformer-based vision models for food quality detection can significantly improve the accuracy and timeliness of assessments, ensuring greater efficiency and safety in food production.

The adaptive local spatiotemporal Vision Transformer represents a novel, high-precision food image classification method developed to address the challenge of distinguishing foods with similar shapes but different nutritional values (Gao et al., 2024a). This method improves classification accuracy through data augmentation and feature enhancement, leveraging the capacity of transformer-based vision models to

support dietary management and health improvement. Researchers incorporated Augmentplus, LayerScale, and multilayer perceptron mechanisms to enhance the local representation of features. The model was trained and validated on the public datasets Food-101 and Vireo Food-172, achieving validation accuracies of 95.17% and 94.29%, respectively. Compared with other advanced self-supervised methods, this approach demonstrated superior performance in food image classification tasks.

In the context of animal product quality assessment, classification and category differentiation are essential for boosting production efficiency and market competitiveness. For instance, in chicken processing, researchers developed an optimised method based on the Swin Transformer to enable real-time classification and detection of different chicken parts by quality (Peng et al., 2024). This method leverages Vision Transformers to capture comprehensive high-level semantic information from chicken part images. By effectively modelling the relationship between local and global features, the approach enhances both classification and detection accuracy. Experimental comparisons showed that the Swin Transformer outperformed YOLOv3-Darknet53, YOLOv3-MobileNetv3, SSD-MobileNetv3, and SSD-VGG16. It achieved higher mean average precision (mAP) scores – exceeding them by 1.62%, 2.13%, 5.26%, and 4.48%, respectively – while also reducing detection time by 16.18 ms, 5.08 ms, 9.38 ms, and 23.48 ms, respectively.

The Swin Transformer not only outperforms existing detection methods in terms of accuracy but also enhances detection speed, demonstrating robustness and suitability for real-time production line applications. Building on this, Gao et al. applied the Swin Transformer for the quantitative detection of carrageenan in beef (Gao et al., 2024b). Compared with biochemical analysis, these transformer-based methods offer a viable alternative for the rapid assessment of quality indicators in animal products.

### **Crop monitoring**

To increase crop yield and land utilisation, modern agriculture often employs mixed cropping methods, where multiple crops are cultivated on the same land to maximise soil nutrient use. However, traditional crop classification and monitoring methods face challenges

in this context due to the varying growth cycles of different crops (Hassan et al., 2021).

In 2020, Z. Li et al. proposed a CNN–Transformer model that integrates multispectral data from various sensors across spatial and spectral scales. The model extracts consistent features and positional information using a multi-layer encoder, followed by classification through a feed-forward and Softmax layer. Experiments on 65 multi-temporal samples from Sentinel-2 A/B and Landsat-8 demonstrated superior performance compared with traditional methods in terms of classification accuracy.

Transformer-based vision models have shown excellent performance in image segmentation, but their application in the food and agriculture fields remains limited. To address this gap, researchers introduced the Agricultural Aerial Transformer (AAFormer) (Shen et al., 2022), designed for the semantic segmentation of aerial farmland images. The architecture combines a Mix Transformer in the encoding phase with a Squeeze-and-Excitation (SE) module in the decoding phase, thereby enhancing anomaly detection. On a validation set, AAFormer achieved a mean intersection over union (mIoU) of 45.44%, demonstrating effectiveness in detecting issues such as drought and nutrient deficiencies.

In large-scale crop monitoring, remote sensing and Earth observation data present further challenges due to the difficulty of obtaining accurately labelled datasets. To address this issue, Wu et al. (2023) applied MoCo-V2 within the Swin Transformer framework to generate an enhanced agricultural dataset. These advancements enable more effective detection of critical agricultural patterns, providing timely alerts to farmers and supporting better-informed decision-making.

### **Disease identification**

Timely screening of plant diseases and analysis of corresponding solutions can optimise resource allocation and enable the rational use of water, fertilisers, and pesticides, thereby reducing production costs and minimising environmental impacts. However, relying solely on manual inspection for disease detection often results in missed opportunities for timely intervention. With the advancement of transformer-based vision models (including ViTs and their variants), disease detection has become more accurate and efficient.

Researchers are continuously seeking ways to extract disease features more efficiently to achieve high-precision classification. In November 2022, Wang et al. (2022b) enhanced a global ViT-based architecture to improve disease diagnosis for tomato plant images. Their dataset, collected from three tomato cultivation areas using drones and cameras, demonstrated an average recognition accuracy of 96.30%. This precise and efficient approach for feature extraction significantly outperforms traditional manual identification methods.

Similarly, Wang et al. (2022a) proposed an enhanced Swin Transformer backbone for cucumber leaf disease recognition. Their approach employed a step-wise patch embedding method to improve feature extraction without adding parameters. Combined with a Grad-CAM-based leaf extraction module, their system – STA-GAN – generated lesions in complex backgrounds to augment datasets, substantially improving disease recognition and data augmentation. Likewise, Parez et al. (2023) introduced Green ViT, a ViT variant for automatic plant disease detection. By partitioning images into smaller patches, Green ViT overcame CNN limitations, outperforming state-of-the-art CNN models in disease detection, thereby reducing costs and improving efficiency.

Not all crop diseases are characterised by visible symptoms; some are extremely subtle. To detect these, Salamai et al. (2023) developed a lesion-aware Vision Transformer targeting subtle variations in rice diseases. Their model incorporated a multi-scale context feature extraction network to capture features across different scales and channels, a weakly supervised rice disease lesion localisation unit to highlight critical lesion areas, and a feature refinement unit to strengthen the relationship between global and local latent spaces. Together, these components enhanced the spatial communication of visual semantics. The system achieved an average accuracy of 98.74% and an F1 score of 98.18% on a public rice disease dataset, demonstrating its robustness in detecting subtle diseases.

The integration of Transformer-based vision models (e.g., Deep Vision Transformer, DVT) with explainable AI has further promoted the adoption of smart agriculture. For instance, Kamal et al. (2024) proposed Deep Vision Transformer (DVT) within an

explainable AI framework, using the “Plant Village” dataset as a case study. Their research focused on two major crops – potatoes and tomatoes – covering six tomato diseases and three potato diseases. The DVT achieved accuracies of 93.56% for tomatoes and 99.95% for potatoes. Importantly, the model visualised its decision-making process, providing farmers with a transparent and interpretable mechanism for early disease detection and management.

### Weed identification

Crop–weed competition is a major factor affecting agricultural growth and yield. Jiang et al. (2022a) developed a deep learning–based semantic segmentation model for weed detection, using data augmentation to enhance the performance of transformer-based models. They compared three architectures – Swin Transformer, SegFormer, and Segmenter – and found that SegFormer achieved the best results (mAcc: 75.18%, mIoU: 65.74%) with only 3.7 million parameters. Silva et al. (2023) proposed a hybrid model combining CNNs and Vision Transformers for soybean weed segmentation. By extending the DeepWeeds dataset with segmentation labels, their model achieved competitive performance compared with state-of-the-art Vision Transformers, while using fewer layers. This demonstrated a lightweight structure suitable for large-scale monitoring.

Weed identification requires not only separating crops and weeds at the same level but also recognising their spatial distribution. To address this, Sun et al. (2024) introduced a Vision Transformer-based neural network for crop and weed identification in sugarcane fields, where crops and weeds often intermingle. Their model integrates multi-scale feature extraction and residual attention layers, achieving 96.97% segmentation accuracy, reducing training parameters by 25%, and showing strong generalisation on the BoniRob dataset – making it a practical solution for mechanical weeding.

### Livestock monitoring

With the development of agricultural automation, commercial free-range farming – particularly poultry farming – faces challenges such as complex backgrounds, multi-scale targets, occlusions, and posture variations. To improve detection accuracy in such environments,

researchers proposed an Efficient Multi-Scale Chicken Detection Model (EMSC-DETR) based on the Real-Time DETR (RT-DETR) framework (X. Li et al., 2024). The model introduces a Space-to-Depth Transformer Module to enhance interactions between local and global features, improving computational efficiency. To address occlusion, it incorporates a Context Transformer, which leverages contextual information between adjacent targets to improve the localisation of overlapping chickens.

Livestock monitoring also requires tracking animal behaviours to prevent the early spread of infectious diseases on farms, ensuring both food safety and human health. Tangirala et al. (2021) developed an end-to-end monitoring system for group-housed pigs, capable of performing instance-level segmentation, tracking, action recognition, and re-identification (STAR) tasks simultaneously. The system employs a transformer-based architecture, “Starformer”, the first multi-object livestock monitoring framework, which learns instance-level embeddings to track pig populations. For benchmarking, the authors introduced the Pigtrace dataset, a curated dataset containing video sequences with instance-level bounding boxes, segmentation, tracking, and activity classification of pigs in real indoor farm environments. Results showed that by optimising the STAR tasks jointly, Starformer outperformed popular baseline models trained on individual tasks.

In large-scale pig farming, manual measurements are often required to assess body size and health, which increases labour costs. To address this, Lu et al. (2022) applied the Swin Transformer to process surveillance data for intelligent identification and segmentation of pigs, enabling contactless monitoring. The model achieved an identification accuracy of 93.0% and a segmentation accuracy of 86.9%. Even under challenging conditions involving overlaps, occlusions, and deformations, it performed robustly. This approach reduces labour costs and advances intelligent, unmanned pig farming, supporting the modernisation of livestock management.

For clarity, Table 1 summarises the different types of transformer-based vision models (including Vision Transformers and their variants) applied in real-world food and agricultural scenarios, along with reported performance.

**Table 1.** Detection efficiency of vision transformers and transformer-based variants in food and agricultural scenarios

Scenarios	Implemented models	Performance (%)
Fruit quality assessment	PD-SegNet (Zhu et al., 2023)	Acc.: 98.89
	Swin Transformer (Xiao et al., 2023)	Acc.: 89.30
	Pre-trained ViT (Knott et al., 2023)	Acc.: 90.00
Animal products monitoring	AlsmViT (Gao et al., 2024a)	Acc.: 95.17
	Swin Transformer (Peng et al., 2024)	Acc.: 97.21
	Swin Transformer (Gao et al., 2024b)	Acc.: 97.50
Crop monitoring	CNN-Transformer (Li et al., 2020)	Acc.: 98.84
	AAFormer (Shen et al., 2022)	mIoU: 45.44
	Swin Transformer (Wu et al., 2023)	mIoU: 43.33
Disease identification	ViT-based architecture (Wang et al., 2022b)	Acc.: 96.30
	STA-GAN (Wang et al., 2022a)	Acc.: 98.97
	Green ViT (Parez et al., 2023)	Acc.: 99.00
	Lesion-aware visual Transformer (Salamai et al., 2023)	Acc.: 98.74
	DVTXAI (Kamal et al., 2024)	Acc.: 99.95
Weed identification	SegFormer (Jiang et al., 2022a)	Acc.: 75.18
	VT-Net (Silva et al., 2023)	Acc.: 93.89
	Visual Transformer variants (Sun et al., 2024)	Acc.: 96.97
Livestock monitoring	EMSC-DETR (Li et al., 2024a)	Acc.: 98.60
	StarFormer (Tangirala et al., 2021)	None
	Swin Transformer (Lu et al., 2022)	Acc.: 93.00

Acc. – Accuracy; mIoU – mean Intersection over Union.

## COMPARATIVE BENCHMARK OF CNN AND ViT ARCHITECTURES

To rigorously evaluate the performance of Vision Transformers (ViTs) and their variants on food- and agriculture-oriented visual tasks, we curated a diverse set of widely used datasets that vary in image size and taxonomic granularity. Using ViT baselines and CNN baselines under identical training protocols, we conducted comprehensive comparisons (Table 2). The results show that ViT variants consistently outperform their CNN counterparts across six food- and agriculture-oriented datasets. Specifically, ViTs achieves higher Top-1 accuracy (or mAP@50 for detection tasks) in every case, with absolute improvements ranging from 3.10 percentage points (pp) on PlantVillage to 24.8 pp on Vireo Food-172. These gains are especially evident in fine-grained and highly multi-class scenarios (i.e., > 100 classes or subtle inter-class differences), where ViTs improve performance by more than 4 pp. This underscores its superior discriminative

power under variable object scales and subtle visual differences common in agricultural imagery. Even when computational budgets are similar ( $\approx 22$ – $26$  M parameters and 4–5 GFLOPs), ViTs achieves an average 8 pp absolute improvement over CNN baselines (e.g., ResNet-50), offering a better accuracy–efficiency trade-off.

These benchmarking findings align with application-specific studies in the food and agriculture domain. For example, models based on ViTs have set new performance benchmarks: FoodCSWin achieved 97.3% accuracy in dietary assessment by optimising the attention mechanism (Xiao et al., 2025a); FGFoodNet enabled ingredient-level fine-grained discrimination (e.g., distinguishing between beef from pork dumplings) by combining CNN-based local feature extraction with transformer-based global modelling (Xiao et al., 2025b); and a fine-grained recognition model based on the Swin Transformer reached 96.5% precision in food quality grading, leveraging its hierarchical architecture (Xiao et al., 2024a). Together, these

**Table 2.** Benchmark of CNN and Vision Transformer (ViT) baselines on common food and agriculture image datasets

Datasets	Resolution	CNN Baselines	ViT Baselines	Params(M)	GFLOPs (G)
FoodX-251 (Xiao et al., 2025a)	224×224	72.23% Top-1 (ResNet50)	76.42% Top-1 (DeiT-s)	26.07/22.15	4.1/4.6
UECFood-256 (Xiao et al., 2025b)	224×224	73.98% Top-1 (TResNet-XL)	77.99% Top-1 (Twins-B)	54.60/55.95	8.6/–
Food-101 (Xiao et al., 2024a)	224×224	n/a	86.14% Top-1 (DeiT-S)	--/21.6	--/4.2
Vireo Food-172 (Li et al., 2025)	224×224	42.9% Top-1 (ResNet50)	67.7% Top-1 (DeiT-B)	58.21/86.57	n/a
PlantDoc (Lin et al., 2025b)	600 × 600	66.95% Top-1 (VGG16)	75.42% Top-1 (ViT-B)	n/a	n/a
PlantVillage (Srivastava et al., 2025)	224×224	90.10% Top-1 (ResNet)	93.20% Top-1 (ViT)	n/a	n/a

results confirm that Vision Transformers and their variants provide stronger discriminative capabilities in food and agricultural scenarios, particularly those characterised by subtle inter-class differences and variable object scales.

## CHALLENGES

Despite the significant potential of transformer-based vision models in food and agricultural object detection, several challenges remain. Below, we outline several key issues that must be addressed in future research and applications.

### Computational complexity

The self-attention mechanism at the core of transformer-based vision models computes relationships between all elements in the input sequence, resulting in quadratic computational complexity, where computation increases quadratically with the length of the input sequence. This issue is particularly pronounced when processing images, which are typically large and require high resolution to capture fine details such as plant diseases, fruits, leaves, and animal characteristics accurately (Saranya et al., 2024). Food image detection faces similar challenges, as high-resolution images are often necessary to identify details such as quality, ripeness, and defects accurately (Jamil et

al., 2023). Reducing computational complexity while preserving accuracy remains a critical challenge for Vision Transformers in agricultural and food-related contexts.

### Large number of model parameters

Transformer-based vision models typically consist of a substantial number of parameters, rendering them more complex than traditional CNN frameworks such as YOLO and Faster R-CNN. This complexity demands greater computational resources, particularly when working with large datasets – for example, those used in remote sensing applications, including drone and satellite imagery, as well as extensive food image datasets like the Open Images Dataset and Tasty 101. These challenges are especially pronounced during both training and inference.

### Environmental variability

Detection in real-world agricultural environments is affected by dynamic factors such as weather, lighting, soil conditions, background clutter, and varying camera angles. These external influences can lead to unstable or inconsistent detection results (Zhu et al., 2020b). Current models often lack robustness in such scenarios, highlighting the need for architectures better adapted to environmental variability (Li et al., 2024a).

### **Small object detection**

Small object detection plays a vital role in agriculture, with applications such as identifying food defects, assessing ripeness, detecting pests and diseases, recognising early crop lesions, and analysing specific livestock features. Because these objects occupy only a few pixels in high-resolution images, traditional CNNs often struggle, particularly in complex or cluttered backgrounds. Transformer-based vision models, with their capacity to model long-range dependencies, offer theoretical advantages in capturing small object features. Nevertheless, they still face significant challenges in extracting fine-grained details, including edge information, multi-scale feature fusion, and hierarchical attention. These limitations restrict the effectiveness of Vision Transformers in small object detection, underscoring the need for continued research and methodological innovation (Min et al., 2023; Sangam et al., 2023; Sivapriya and Suresh, 2023).

### **Over-reliance on large-scale labelled data**

Transformer-based vision models (e.g., ViT, Swin Transformer) achieve state-of-the-art accuracy when pretrained on large-scale datasets such as ImageNet. However, in agricultural domains – where annotated data is often scarce (e.g., rare crop diseases or diverse orchard scenes) – their generalisation ability degrades under low-data regimes. Lacking locality-centric inductive biases, these models often underperform compact CNNs (e.g., YOLOv11) in few-shot tasks, where the latter’s built-in translational equivariance and parameter efficiency confer superior robustness.

### **Poor interpretability**

The opaque nature of attention mechanisms in transformer-based vision models makes it difficult to explain model decisions at the pixel-level – for example, why a fruit is classified as defective. Even state-of-the-art post-hoc saliency methods (e.g., DVT) struggle to provide explanations with sufficient fidelity for safety-critical food-inspection systems. This lack of interpretability undermines trust among stakeholders and complicates regulatory approval.

## **FUTURE PROSPECTS**

Drawing on the development history, current advancements, and diverse applications of transformer-based vision models in object detection – as well as the challenges they face – we outline several future research directions and opportunities. For reference, Table 3 summarises representative transformer-based vision models for object detection and their respective improvements.

### **Multimodal fusion**

Vision Transformers show strong potential for multimodal fusion in object detection. Leveraging self-attention, they can integrate data from diverse sources, including visual images, spectral information, meteorological data, and sensor readings. This cross-modal feature alignment enables the model to capture both global and local features in complex environments, thereby improving detection accuracy and robustness. During the fusion process, Vision Transformers can dynamically adjust the weights of different modalities, prioritising the most relevant inputs for specific tasks and enhancing decision-making accuracy. Furthermore, they can mitigate the impact of incomplete or noisy data through complementary learning mechanisms, inferring missing information from other modalities. This multimodal fusion capability applies not only to tasks such as pest detection, precision irrigation, and fertilisation but also to crop growth prediction, offering more comprehensive and reliable decision support for smart agriculture (Kalamkar and Amalanathan, 2024).

### **Reducing computational costs**

The high computational costs of transformer-based vision models limit their widespread adoption in practical applications. Future object detection models based on Vision Transformers may incorporate optimisation strategies such as model compression, pruning, and dynamic adjustment to reduce unnecessary computational overhead while maintaining detection accuracy. These approaches can effectively lower floating-point operations (FLOPs), inference time, and energy consumption. Furthermore, techniques such as sequence parallelism and selective activation recomputation provide promising directions for further reducing the computational costs of Vision Transformers (Fan et al., 2024; Korthikanti et al., 2023).

### **Interpretability and transparency**

Because food and agricultural production is inherently uncertain, ensuring model interpretability and transparency remains a critical research priority. Explainable systems should clarify how machine vision evaluates product quality based on specific features. These insights can reveal both strengths and weaknesses of food and agricultural products and guide producers in improving quality. Despite their strong performance, transformer-based vision models are difficult to interpret due to their complex attention mechanisms. Future work is therefore likely to focus on developing interpretable models that allow consumers, regulators, manufacturers, farmers, and experts to better understand model decision-making, thereby strengthening trust and confidence in their outputs (Feng and Xu, 2024; Kamal et al., 2024).

### **Integration with the IoT and edge computing**

With the widespread application of IoT technologies in food and agriculture, Vision Transformers are expected to become increasingly integrated with smart devices. For example, drones and intelligent sensors can capture farmland images, while edge computing enables lightweight transformer-based vision models to run on resource-constrained devices, supporting real-time data analysis. Future detection systems are likely to incorporate real-time learning and adaptive processing mechanisms. By responding dynamically to changing conditions and continuously improving through onsite feedback, these models can optimise performance over time and advance the broader application of Vision Transformers in agriculture (Zhang and Lv, 2024).

### **Small-sample learning for niche scenarios**

In smart agriculture, tasks such as disease identification and growth monitoring of rare crops (e.g., wolfberry, *Panax notoginseng*) are often hindered by data scarcity. Factors such as limited planting scope, difficulties in sample collection, and the low incidence of disease cases prevent the accumulation of sufficient annotated data. As a result, transformer-based vision models trained on large-scale datasets perform poorly when applied in practice. Few-shot learning has emerged as a promising solution to this challenge. The integration of adaptive feedback cross-loop self-KD

methods with meta-learning frameworks and data augmentation techniques offers a feasible pathway for adapting models to rare-crop scenarios (Lin et al., 2025a).

### **Lightweight models for edge devices**

Conventional Vision Transformers are often too large for real-time deployment on agricultural edge devices. To address this limitation, techniques such as pruning, knowledge distillation, and CNN–Vision Transformer hybrid architectures can be applied synergistically (Yu et al., 2022). In this approach, redundant attention heads and channels are structurally removed, while a pruned student network learns essential representations under the guidance of a high-precision teacher. Subsequent quantisation and compilation optimisations further reduce the model’s footprint. As a result, network size can be significantly reduced without substantial accuracy loss, enabling efficient execution on low-cost, resource-constrained sensors or drones and supporting genuine real-time monitoring and decision-making in the field.

In summary, the evolution of transformer-based vision models in object detection demonstrates that, despite persistent challenges, the approach retains substantial potential for further development. At the technical level, issues such as high computational demands, model complexity, and dependence on large-scale datasets necessitate ongoing innovation to improve both efficiency and practicality. Data-related challenges highlight the critical role of high-quality annotated datasets and the need for timely updates, emphasising the importance of more effective data collection strategies and the establishment of open data-sharing platforms. From an application standpoint, factors including deployment environment variability, user acceptance, and training requirements remain decisive in shaping the broader adoption of this technology.

## **CONCLUSIONS**

CNN-based models, such as R-CNN, Fast R-CNN, SSD, and YOLO (and their variants), have long dominated object detection tasks across multiple domains. In recent years, however, transformer-based models – particularly the Vision Transformer (ViT) and

**Table 3.** Transformer-based vision models and their improvements in object detection

Variant types	Improvements
DETR (Carion et al., 2020)	<ol style="list-style-type: none"> <li>(1) End-to-End Training: DETR proposes an end-to-end framework that eliminates the need for hand-crafted components used in traditional object detection methods.</li> <li>(2) Hungarian Algorithm for Matching: Employs the Hungarian algorithm to match predicted boxes with ground-truth boxes, providing a more robust assignment mechanism.</li> <li>(3) Global Context Modelling: Uses the attention mechanism to model global context effectively, improving detection in complex scenes.</li> <li>(4) Feature Fusion Across Domains: Leverages the Transformer architecture to integrate multi-level features, enhancing the detection of small and occluded objects.</li> </ol>
ViT (Dosovitskiy et al., 2021)	<ol style="list-style-type: none"> <li>(1) Patch-Based Input: Splits images into fixed-size patches, embeds them linearly, and processes them as sequences, similar to NLP tasks.</li> <li>(2) Position Encoding: Introduces learnable position embeddings to retain spatial information.</li> <li>(3) Training Strategy: Achieves strong performance when pre-trained on large datasets, benefiting from larger models and extended training durations.</li> <li>(4) Global Feature Learning: Uses the self-attention mechanism to capture global dependencies effectively, enhancing performance on diverse vision tasks.</li> </ol>
IPT (Chen et al., 2021)	<ol style="list-style-type: none"> <li>(1) Focus on Low-Level Computer Vision Tasks: Targets tasks such as denoising, super-resolution, and deraining.</li> <li>(2) Leveraging Large-Scale Datasets: Utilises the ImageNet benchmark to generate large sets of corrupted image pairs, maximising Transformer capabilities.</li> <li>(3) Multi-Head and Multi-Tail Structure: Employs a multi-head and multi-tail design to strengthen feature learning.</li> <li>(4) Introduction of Contrastive Learning: Incorporates contrastive learning to improve adaptability to diverse image processing tasks.</li> <li>(5) Efficient Fine-Tuning Capability: After pre-training, can be efficiently fine-tuned for specific tasks, outperforming state-of-the-art methods on multiple low-level benchmarks.</li> </ol>
Segformer (Xie et al., 2021)	<ol style="list-style-type: none"> <li>(1) Simple and Efficient Framework: Combines Transformers with lightweight MLP decoders for semantic segmentation.</li> <li>(2) Novel Hierarchical Encoder: Employs a hierarchical Transformer encoder that produces multi-scale features without positional encoding, avoiding performance degradation.</li> <li>(3) Simplified Decoder: Uses an MLP decoder to aggregate multi-layer information, combining local and global attention for stronger feature representation.</li> <li>(4) Efficiency and Performance Gains: Scales from SegFormer-B0 to SegFormer-B5, achieving superior efficiency and accuracy compared to prior methods.</li> <li>(5) Performance Improvement: For example, SegFormer-B4 achieves 50.3% mIoU on ADE20K with 64 million parameters – five times smaller and 2.2% more accurate than the previous best. The top model, SegFormer-B5, reaches 84.0% mIoU on Cityscapes validation and demonstrates excellent zero-shot robustness on Cityscapes-C.</li> </ol>
TNT (Han et al., 2021)	<ol style="list-style-type: none"> <li>(1) Fine-Grained Feature Extraction: Divides local patches into smaller patches (“visual words”), improving capture of features at different scales and positions.</li> <li>(2) Local Attention Mechanism: Calculates attention within local patches to improve performance on complex images.</li> <li>(3) Low-Cost Attention Calculation: Computes attention between visual words at low computational cost, improving efficiency.</li> <li>(4) Enhanced Representation Ability: Aggregates features of both visual words and visual sentences for stronger overall representation.</li> <li>(5) Outstanding Performance: Achieves 81.5% top-1 accuracy on ImageNet and outperforms state-of-the-art visual Transformers with comparable computational costs.</li> </ol>
Swin Transformer (Liu et al., 2021)	<ol style="list-style-type: none"> <li>(1) Hierarchical Architecture: Employs a hierarchical structure for multi-scale feature modelling, enhancing adaptability across vision tasks.</li> <li>(2) Shifted Window Mechanism: Restricts self-attention computation to non-overlapping local windows while enabling cross-window connections, improving efficiency.</li> <li>(3) Linear Computational Complexity: Achieves linear computational complexity with respect to image size, making it more efficient for high-resolution inputs compared to standard Transformers.</li> <li>(4) Broad Applicability: Performs strongly across image classification, object detection, and semantic segmentation, achieving state-of-the-art results in these tasks.</li> <li>(5) Performance Improvement: Outperforms traditional Transformers in accuracy, particularly for object detection and semantic segmentation tasks, underscoring its effectiveness as a visual backbone.</li> </ol>

its derivatives – have gained prominence in computer vision. Their self-attention mechanisms and ability to capture long-range dependencies, combined with increasingly lightweight architectures, enable transformer-based vision models to deliver superior precision and efficiency. These advances have opened new opportunities in food and agriculture, including the detection of chemical and physical properties of food, quality classification, identification of adulteration and counterfeit products, crop monitoring, pest and disease detection, and weed identification. This study provides the first comprehensive review of transformer-based vision models for object detection in food and agriculture. It examines the topic from four perspectives: (1) the evolution of transformer architectures; (2) the architectures and variants of transformer-based models; (3) their applications in real-world agricultural and food-related scenarios; and (4) the challenges and future prospects of transformer-based approaches. Our analysis shows that transformer-based vision models offer more intelligent, sustainable, and efficient decision-support capabilities compared with traditional approaches that rely on manual inspection and chemical reagents. They hold significant promise for researchers, farmers, producers, and consumers in advancing precision, automation, and sustainability in food and agriculture systems.

## FUNDING INFORMATION

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.

## DECLARATIONS

### Data statement

All data supporting this study has been included in this manuscript.

## Ethical Approval

Not applicable.

## Competing Interests

The authors declare that they have no conflicts of interest.

## OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

## REFERENCES

- Abbaspour-Gilandeh, Y., Aghabara, A., Davari, M., Maja, J. M. (2022). Feasibility of using computer vision and artificial intelligence techniques in detection of some apple pests and diseases. *Applied. Sciences*, 12(2), 906. <https://doi.org/10.3390/app12020906>
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ..., Zoph, B. (2023). GPT-4 technical report. arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>
- Ale, L., Sheta, A., Li, L., Wang, Y., Zhang, N. (2019). Deep learning based plant disease detection for smart agriculture. In: *IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA.
- Aradhya, H. R. (2019). Object detection and tracking using deep learning and artificial intelligence for video surveillance applications. *Int. J. Adv. Comp. Sci. Appl.*, 10, (12). <https://doi.org/10.14569/IJACSA.2019.0101269>
- Ariza-Sentís, M., Vélez, S., Martínez-Peña, R., Baja, H., Valente, J. (2024). Object detection and tracking in precision farming: A systematic review. *Comp.*

- Electr. Agric.*, 219, 108757. <https://doi.org/10.1016/j.compag.2024.108757>
- Bianco, S., Buzzelli, M., Chiriaco, G., Napoletano, P., Piccoli, F. (2023). Food recognition with visual transformers. In: 2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany.
- Bochkovskiy, A. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv:2004.10934*. <https://doi.org/10.48550/arXiv.2004.10934>
- Cai, J., Tao, J., Ma, Y., Fan, X., Cheng, L. (2020). Fruit image recognition and classification method based on improved single shot multi-box detector. *J. Physic. Conf. Ser.*, 1629(1), 012010. <https://doi.org/10.1088/1742-6596/1629/1/012010>
- Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., ..., Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 2814–2830. <https://doi.org/10.1109/TKDE.2024.3361474>
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., ..., Zagoruyko, S. (2020). End-to-end object detection with transformers. In: *Proceedings of the European Conference on Computer Vision (ECCV 2020)*, Glasgow, UK. <https://doi.org/10.48550/arXiv.2005.12872>
- Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., ..., Gao, W. (2021). Pre-trained image processing transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA. <https://doi.org/10.48550/arXiv.2012.00364>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., ..., Hounsby, N. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2010.11929>
- Fahad, L. G., Tahir, S. F., Rasheed, U., Saqib, H., Hassan, M., et al. (2022). Fruits and vegetables freshness categorization using deep learning. *Comp. Mater. Contin.*, 71(3). <https://doi.org/10.32604/cmc.2022.023357>
- Fan, Q., Huang, H., Zhou, X., He, R. (2023). Lightweight vision transformer with bidirectional interaction. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA. <https://doi.org/10.48550/arXiv.2306.00396>
- Feng, J., Xu, X. (2024). Deciphering plant seedlings: Enhancing classification and interpretability with vision transformers. In: *Proceedings of the 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL, 2024)*, Zhuhai, China. <https://doi.org/10.1109/CVIDL62147.2024.10604151>
- Gao, X., Xiao, Z., Deng, Z. (2024a). High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *J. Food Eng.*, 365, 111833. <https://doi.org/10.1016/j.jfoodeng.2023.111833>
- Gao, Z., Chen, S., Huang, J., Cai, H. (2024b). Real-time quantitative detection of hydrocolloid adulteration in meat based on Swin Transformer and smartphone. *J. Food Sci.*, 89(7), 4359–4371. <https://doi.org/10.1111/1750-3841.17159>
- Gao, Z., Huang, J., Chen, J., Shao, T., Ni, H., Cai, H. (2024c). Deep transfer learning-based computer vision for real-time harvest period classification and impurity detection of *Porphyra haitnensis*. *Aquacult. Int.*, 32, 5171–5198. <https://doi.org/10.1007/s10499-024-01422-6>
- Ghazal, S., Munir, A., Qureshi, W. S. (2024). Computer vision in smart agriculture and precision farming: Techniques and applications. *Artif. Intell. Agric.*, 13, 64–83. <https://doi.org/10.1016/j.aiaa.2024.06.004>
- Gong, X., Zhang, S. (2023). A high-precision detection method of apple leaf diseases using improved Faster R-CNN. *Agriculture*, 13(2), 240. <https://doi.org/10.3390/agriculture13020240>
- Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y. (2021). Transformer in Transformer. In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2021)*, Red Hook, NY, USA. <https://doi.org/10.48550/arXiv.2103.00112>
- Harnsoongnoen, S., Jaroensuk, N. (2021). The grades and freshness assessment of eggs based on density detection using machine vision and weighing sensor. *Sci. Rep.*, 11(1), 16640. <https://doi.org/10.1038/s41598-021-96140-x>
- Hassan, S. I., Alam, M. M., Illahi, U., Al Ghamdi, M. A., Almotiri, S. H., et al. (2021). A systematic review on monitoring and advanced control strategies in smart agriculture. *IEEE Access*, 9, 32517–32548. <https://doi.org/10.1109/ACCESS.2021.3057865>
- He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy. <https://doi.org/10.48550/arXiv.1703.06870>
- Jabir, B., Moutaouakil, K. E., Falih, N. (2023). Developing an efficient system with Mask R-CNN for agricultural applications. *Agris on-line Papers in Economics and Informatics*, 15(1), 61–72. <https://doi.org/10.22004/ag.econ.334659>

- Jamil, S., Jalil Piran, M., Kwon, O. J. (2023). A comprehensive survey of transformers for computer vision. *Drones*, 7(5), 287. <https://doi.org/10.3390/drones7050287>
- Jiang, K., Afzaal, U., Lee, J. (2022a). Transformer-based weed segmentation for grass management. *Sensors*, 23(1), 65. <https://doi.org/10.3390/s23010065>
- Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B. (2022b). A Review of Yolo algorithm developments. *Proced. Comp. Sci.*, 199, 1066–1073. <https://doi.org/10.1016/j.procs.2022.01.135>
- Kalamkar, S., Amalanathan, G. M. (2024). MDA-ViT: Multimodal image fusion using dual attention vision transformer. *Multimedia Tools and Applications*. Advance online publication. <https://doi.org/10.1007/s11042-024-19968-1>
- Kamal, S., Sharma, P., Gupta, P., Siddiqui, M. K., Dutt, A., Singh, A. (2024). DVTXAI: A novel deep Vision Transformer with an explainable AI-based framework and its application in agriculture. *J. Supercom.*, 81. <https://doi.org/10.21203/rs.3.rs-4752298/v1>
- Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., ..., Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10s), 1–41. <https://doi.org/10.1145/3505244>
- Khan, A., Hassan, T., Shafay, M., Fahmy, I., Werghi, N., ..., Hussain, I. (2023). Tomato maturity recognition with convolutional transformers. *Sci. Rep.*, 13(1), 22885. <https://doi.org/10.1038/s41598-023-50129-w>
- Khanam, R., Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv:2410.17725*. <https://doi.org/10.48550/arXiv.2410.17725>
- Knott, M., Perez-Cruz, F., Defraeye, T. (2023). Facilitated machine learning for image-based fruit quality assessment. *J. Food Eng.*, 345, 111401. <https://doi.org/10.1016/j.jfoodeng.2022.111401>
- Korthikanti, V. A., Casper, J., Lym, S., McAfee, L., Andersch, M., ..., Catanzaro, B. (2023). Reducing activation recomputation in large transformer models. In: *Proceedings of Machine Learning and Systems*, 5, 341–353. <https://doi.org/10.48550/arXiv.2205.05198>
- Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Devlin, J., Chang, M. W., Lee, J., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, North American. Retrieved from: <https://arxiv.org/pdf/1810.04805>
- Lee, J., Toutanova, K. (2019). Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota. <https://doi.org/10.18653/v1/N19-1423>
- Li, C., Sun, X., Cai, J. (2019). Intelligent mobile drone system based on real-time object detection. *J. Artif. Intell.*, 1(1), 1–8. <https://doi.org/10.32604/jai.2019.06064>
- Li, Z., Chen, G., Zhang, T. (2020). A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. *IEEE J. Select. Topic. Appl. Earth Observ. Remot. Sens.*, 13, 847–858. <https://doi.org/10.1109/JSTARS.2020.2971763>
- Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., ..., Yan, J. (2021a). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*. <https://doi.org/10.48550/arXiv.2110.05208>
- Li, L., Zhang, S., Wang, B. (2021b). Plant disease detection and classification by deep learning – a review. *IEEE Access*, 9, 56683–56698. <https://doi.org/10.1109/ACCESS.2021.3069646>
- Li, X., Cai, M., Tan, X., Yin, C., Chen, W., ..., Han, Y. (2024a). An efficient transformer network for detecting multi-scale chicken in complex free-range farming environments via improved RT-DETR. *Comp. Elect. Agric.*, 224, 109160. <https://doi.org/10.1016/j.compag.2024.109160>
- Li, Z., Dong, Y., Shen, L., Liu, Y., Pei, Y., ..., Ma, J. (2024b). Development and challenges of object detection: A survey. *Neurocomputing*, 598, 128102. <https://doi.org/10.1016/j.neucom.2024.128102>
- Li, J., Xu, H., Zhu, X., Xiong, J., & Zhang, X. (2025). FSF-ViT: Image augmentation and adaptive global-local feature fusion for Few-Shot Food classification. *Food Chem.*, 492(3), 145276. <https://doi.org/10.1016/j.foodchem.2025.145276>
- Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy. <https://doi.org/10.48550/arXiv.1708.02002>
- Lin, Y., Cai, Y., Chen, H., Cai, Y., Lin, Z., ..., Ni, H. (2025a). Adaptive feedback cross-loop for preserving and robust spectral information optimization without spectral processing in few-shot learning. *Measur. Sci. Technol.*, 36(7). <https://doi.org/10.1088/1361-6501/ad2a>

- Lin, J., Chen, X., Lou, L., You, L., Cernava, T., ..., Zhang, X. (2025b). DIEC-ViT: Discriminative information enhanced contrastive vision Transformer for the identification of plant diseases in complex environments. *Expert Syst. Appl.*, 281, 127730. <https://doi.org/10.1016/j.eswa.2025.127730>
- Lipton, Z. C., Berkowitz, J., Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019*. <https://doi.org/10.48550/arXiv.1506.00019>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., ..., Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada.
- Lu, J., Wang, W., Zhao, K., Wang, H. (2022). Recognition and segmentation of individual pigs based on Swin Transformer. *Anim. Gen.*, 53(6), 794–802. <https://doi.org/10.1111/age.13259>
- Marinoudi, V., Sørensen, C. G., Pearson, S., Bochtis, D. (2019). Robotics and labour in agriculture. A context consideration. *Biosyst. Eng.*, 184, 111–121. <https://doi.org/10.1016/j.biosystemseng.2019.06.013>
- Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., ..., Wang, J. (2021). Conditional DETR for fast training convergence. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada. <https://doi.org/10.48550/arXiv.2108.06152>
- Min, L., Fan, Z., Lv, Q., Reda, M., Shen, L., Wang, B. (2023). Yolo-dcti: Small object detection in remote sensing base on contextual transformer enhancement. *Remote Sens.*, 15(16), 3970. <https://doi.org/10.3390/rs15163970>
- Mostafa, S. A., Ravi, S., Zebari, D. A., Zebari, N. A., Mohammed, M. A., ..., Ding, W., (2024). A YOLO-based deep learning model for real-time face mask detection via drone surveillance in public spaces. *Inf. Sci.*, 676, 120865. <https://doi.org/10.1016/j.ins.2024.120865>
- Mu, Y., Feng, R., Ni, R., Li, J., Luo, T., ..., Hu, T. (2022). A faster R-CNN-based model for the identification of weed seedling. *Agronomy*, 12(11), 2867. <https://doi.org/10.3390/agronomy12112867>
- Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sens.*, 10(8), 1217. <https://doi.org/10.3390/rs10081217>
- Ning, C., Zhou, H., Song, Y., Tang, J. (2017). Inception single shot multibox detector for object detection. In: *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Hong Kong, China. <https://doi.org/10.1109/ICMEW.2017.8026312>
- Parez, S., Dilshad, N., Alghamdi, N. S., Alanazi, T. M., Lee, J. W. (2023). Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. *Sensors*, 23(15), 6949. <https://doi.org/10.3390/s23156949>
- Patrício, D. I., Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Comput. Electr. Agric.*, 153, 69–81. <https://doi.org/10.1016/j.compag.2018.08.001>
- Peng, X., Xu, C., Zhang, P., Fu, D., Chen, Y., Hu, Z. (2024). Computer vision classification detection of chicken parts based on optimized Swin-Transformer. *CyTA – Journal of Food*, 22(1), 2347480. <https://doi.org/10.1080/19476337.2024.2347480>
- Rajamohanam, R., Latha, B. C. (2023). An optimized YOLO v5 model for tomato leaf disease classification with field dataset. *Eng. Technol. Appl. Sci. Res.*, 13(6), 12033–12038. <https://doi.org/10.48084/etasr.6377>
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA. <https://doi.org/10.48550/arXiv.1506.02640>
- Redmon, J. (2018). Yolov3: An incremental improvement. *arXiv:1804.02767*, [abs/1804.02767](https://doi.org/10.48550/arXiv.1804.02767). <https://doi.org/10.48550/arXiv.1804.02767>
- Redmon, J., Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA. <https://doi.org/10.48550/arXiv.1612.08242>
- Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Salamai, A. A., Ajabnoor, N., Khalid, W. E., Ali, M. M., Murayr, A. A. (2023). Lesion-aware visual transformer network for paddy diseases detection in precision agriculture. *Eur. J. Agro.*, 148, 126884. <https://doi.org/10.1016/j.eja.2023.126884>
- Sangam, T., Dave, I. R., Sultani, W., Shah, M. (2023). Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom. <https://doi.org/10.1109/ICRA48891.2023.10161433>

- Saranya, T., Deisy, C., Sridevi, S. (2024). Efficient agricultural pest classification using vision transformer with hybrid pooled multihead attention. *Computers in Biology and Medicine*, 177, 108584. <https://doi.org/10.1016/j.combiomed.2024.108584>
- Sharma, A., Jain, A., Gupta, P., Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9, 4843–4873. <https://doi.org/10.1109/ACCESS.2020.3048415>
- Shen, Y., Wang, L., Jin, Y. (2022). AAFFormer: a multi-modal transformer network for aerial agricultural images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA. <https://doi.org/10.1109/CVPRW56347.2022.00177>
- Shipitko, O., Kibalov, V., Abramov, M. (2020). Linear features observation model for autonomous vehicle localization. In: *Proceedings of the 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Shenzhen, China. <https://doi.org/10.48550/arXiv.2002.12731>
- Silva, L., Drews, P., de Bem, R. (2023). Soybean weeds segmentation using VT-Net: A convolutional-transformer model. In: *Proceedings of the 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Rio Grande, RS, Brazil. <https://doi.org/10.1109/SIBGRAPI59091.2023.10347167>
- Singh, G., Sethi, G. K., Singh, S. (2021). Survey on machine learning and deep learning techniques for agriculture land. *SN Comput. Sci.*, 2(6), 487. <https://doi.org/10.1007/s42979-021-00929-6>
- Sivapriya, M., Suresh, S. (2023). ViT-DexiNet: A vision transformer-based edge detection operator for small object detection in SAR images. *Int. J. Remot. Sens.*, 44(22), 7057–7084. <https://doi.org/10.1080/01431161.2023.2277167>
- Srivastava, M., Sisaudia, V., Meena, J. (2025). AgriTL-ViT: A vision transformer model with attention techniques for classification of plant leaf disease. *Expert Syst. Appl.*, 294, 128793. <https://doi.org/10.1016/j.eswa.2025.128793>
- Strudel, R., Garcia, R., Laptev, I., Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada. <https://doi.org/10.48550/arXiv.2105.05633>
- Sun, C., Zhang, M., Zhou, M., Zhou, X. (2024). An improved transformer network with multi-scale convolution for weed identification in sugarcane field. *IEEE Access*, 12, 31168–31181. <https://doi.org/10.1109/ACCESS.2024.3368911>
- Tangirala, B., Bhandari, I., Laszlo, D., Gupta, D. K., Thomas, R. M., et al. (2021). Livestock monitoring with transformer. In: *Proceedings of the 32nd British Machine Vision Conference (BMVC)*. <https://doi.org/10.48550/arXiv.2111.00801>
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In: *Proceedings of the International Conference on Machine Learning (ICML)*. <https://doi.org/10.48550/arXiv.2012.12877>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., ..., Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, California, USA. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, B. (2022). Automatic mushroom species classification model for foodborne disease prevention based on vision transformer. *J. Food Qual.*, 2022(1), 1173102. <https://doi.org/10.1155/2022/1173102>
- Wang, C., Du, P., Wu, H., Li, J., Zhao, C., Zhu, H. (2021a). A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net. *Computers and Electronics in Agriculture*, 189, 106373. <https://doi.org/10.1016/j.compag.2021.106373>
- Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., ..., Shao, L. (2021b). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada. <https://doi.org/10.48550/arXiv.2102.12122>
- Wang, F., Rao, Y., Luo, Q., Jin, X., Jiang, Z., ..., Li, S. (2022a). Practical cucumber leaf disease recognition using improved Swin Transformer and small sample size. *Computers and Electronics in Agriculture*, 199, 107163. <https://doi.org/10.1016/j.compag.2022.107163>
- Wang, Y., Chen, Y., Wang, D. (2022b). Convolution network enlightened transformer for regional crop disease classification. *Electronics*, 11(19), 3174. <https://doi.org/10.3390/electronics11193174>
- Wu, J., Pichler, D., Marley, D., Wilson, D., Hovakimyan, N., Hobbs, J. (2023). Extended agriculture-vision: An extension of a large aerial image dataset for agricultural pattern analysis. *arXiv:2303.02460*. <https://doi.org/10.48550/arXiv.2303.02460>
- Xiao, B., Nguyen, M., Yan, W. Q. (2023). Fruit ripeness identification using transformers. *Appl. Intell.*, 53(19), 22488–22499. <https://doi.org/10.1007/s10489-023-04799-8>

- Xiao, Z., Diao, G., Deng, Z. (2024a). Fine grained food image recognition based on Swin Transformer. *J. Food Eng.*, 380, 112134. <https://doi.org/10.1016/j.jfood-eng.2024.112134>
- Xiao, Y., Huang, Y., Qiu, J., Cai, H., Ni, H. (2024b). Smartphone-based pH titration for liquid food applications. *Chem. Paper.*, 78(16), 8849–8862. <https://doi.org/10.1007/s11696-024-03715-9>
- Xiao, Z., Ling, R., Deng, Z. (2025a). FoodCSWin: A high-accuracy food image recognition model for dietary assessment. *J. Food Compos. Anal.*, 139, 107110. <https://doi.org/10.1016/j.jfca.2024.107110>
- Xiao, Z., Sun, Y., Deng, Z. (2025b). FGFoodNet: Ingredient-perceived fine-grained food recognition for dietary monitoring. *J. Food Measur. Character.*, 1–17. <https://doi.org/10.1007/s11694-025-03439-8>
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.*, 34, 12077–12090. <https://doi.org/10.48550/arXiv.2105.15203>
- Yao, Z., Ai, J., Li, B., Zhang, C. (2021). Efficient DETR: improving end-to-end object detector with dense prior. *arXiv:2104.01318*. <https://doi.org/10.48550/arXiv.2104.01318>
- Yu, F., Huang, K., Wang, M., Cheng, Y., Chu, W., Cui, L. (2022). Width & depth pruning for vision transformers. In: *Proceedings of the AAAI conference on artificial intelligence (AAAI)*. <https://doi.org/10.1609/aaai.v36i3.20222>
- Zhang, L., Rao, A., Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France. <https://doi.org/10.48550/arXiv.2302.05543>
- Zhang, Q., Liu, Y., Gong, C., Chen, Y., Yu, H. (2020). Applications of deep learning for dense scenes analysis in agriculture: A review. *Sensors*, 20(5), 1520. <https://doi.org/10.3390/s20051520>
- Zhang, Y., Lv, C. (2024). TinySegformer: A lightweight visual segmentation model for real-time agricultural pest detection. *Comp. Elect. Agric.*, 218, 108740. <https://doi.org/10.1016/j.compag.2024.108740>
- Zhao, Z., Zheng, P., Xu, S., Wu, X. (2019). Object detection with deep learning: A review. *IEEE Trans. Neural Net. Learn. Syst.*, 30(11), 3212–3232. <https://doi.org/10.1109/TNNLS.2018.2876865>
- Zhao, S., Hao, G., Zhang, Y., Wang, S. (2021). A real-time classification and detection method for mutton parts based on single shot multi-box detector. *J. Food Process Eng.*, 44(8), e13749. <https://doi.org/10.1111/jfpe.13749>
- Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE*, 109(5), 820–838. <https://doi.org/10.1109/JPROC.2021.3054390>
- Zhu, L., Spachos, P., Pensini, E., Plataniotis, K. N. (2021). Deep learning and machine vision for food processing: A survey. *Curr. Res. Food Sci.*, 4, 233–249. <https://doi.org/10.1016/j.crfs.2021.03.009>
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. (2020a). Deformable DETR: Deformable Transformers for end-to-end object detection. *arXiv:2010.04159*. <https://doi.org/10.48550/arXiv.2010.04159>
- Zhu, Y., Zhao, X., Zhao, C., Wang, J., Lu, H. (2020b). FoodDet: Detecting foods in refrigerator with supervised transformer network. *Neurocomputing*, 379, 162–171. <https://doi.org/10.1016/j.neucom.2019.10.106>
- Zhu, Z., Jiang, M., Dong, J., Wu, S., Ma, F. (2023). PD-SegNet: Semantic segmentation of small agricultural targets in complex environments. *IEEE Access*, 11, 90214–90226. <https://doi.org/10.1109/ACCESS.2023.3284036>

