Journal of Food and Nutrition Research (ISSN 1336-8672) Vol. *65*, 2026, No. 1, pp. 1–12

DOI: https://doi.org/10.64122/HOYB1678

REVIEW

Comparative review of Vision Transformer and YOLO in food and agriculture

LIN MAOLAN – GAO ZHENCHANG – LIAO WENLIANG – CAI HONGHAO

Summary

Computer vision is vital in food and agriculture, with object detection being crucial for automation. Among these, You Only Look Once (YOLO) and Vision Transformer (ViT) models have emerged as two influential approaches. Despite their architectural differences, the two approaches often complement each other in practice. However, direct comparative studies remain limited and the literature fragmented. This paper starts by mapping the road from YOLO version 1 to the latest convolutional neural networks (CNN) and from the original Transformer to recent vision variants, showing why key designs were made and how they set the two streams apart. We benchmark the base architectures and their popular variants on food-and-agriculture datasets and quantify the gap between reported accuracy and reproduced/emerged accuracy. We conclude with an outlook on open challenges, emerging remedies, and recent advances poised to define the next generation of both paradigms. By integrating key literature (peer-reviewed articles within the last 10 years), this study constructs a systematic comparison of YOLO and ViT in the food and agricultural fields, which not only clarifies the technical boundaries and applicable scenarios of the two algorithms but also provides a theoretical basis and practical guidance for algorithm selection and optimisation in actual production.

Keywords

deep learning; convolutional neural network; computer vision; intelligent agriculture; food quality; attention mechanism; You Only Look Once

Against the backdrop of the rapid development of digital technologies in food and agriculture, computer vision has significantly enhanced efficiency in this field. Technologies such as food classification [1], pest and disease diagnosis [2], adulteration identification [3], animal product monitoring [4] and others heavily rely on object detection technology. Currently, mainstream applications in this domain are primarily dominated by two major algorithmic paradigms: the You Only Look Once (YOLO) series, and Vision Transformer (ViT) and its variants.

Different neural network architectures have achieved remarkable progress in various visual tasks. For example, two-stage models like Faster Region-Based Convolutional Neural Network (Fast R-CNN) [5], RetinaNet [6], and Mask Region-Based Convolutional Neural Network (Mask R-CNN) [7], which extends Fast R-CNN; as well as one-stage models like Single Shot MultiBox Detector (SSD) [8]; and others. As a representative of convolutional neural networks (CNN)-based detectors, the YOLO series is known for its end-to-end real-time performance: it transforms object detection into a regression task, simultaneously predicting bounding boxes and class labels within a single network. This makes it particularly suitable for time-sensitive field scenarios, such as real-time fruit counting during harvest or livestock behaviour monitoring [9]. Its derivative versions

Lin Maolan, Liao Wenliang, Cai Honghao, Department of Physics, School of Science, Jimei University, Yinjiang Road 185, 361021 Xiamen, Fujian Province, China.

Gao Zhenchang, School of Information Science and Technology, ShanghaiTech University, Middle Huaxia Road 393, 201210 Shanghai, China; Guangdong Institution of Intelligent Science and Technology, Huandao North Road 2515, 519031 Zhuhai, Guangdong Province, China.

Correspondence author:
Cai Honghao, e-mail: hhcai@jmu.edu.cn
Liao Wenliang, e-mail: 200661000118@jmu.edu.cn

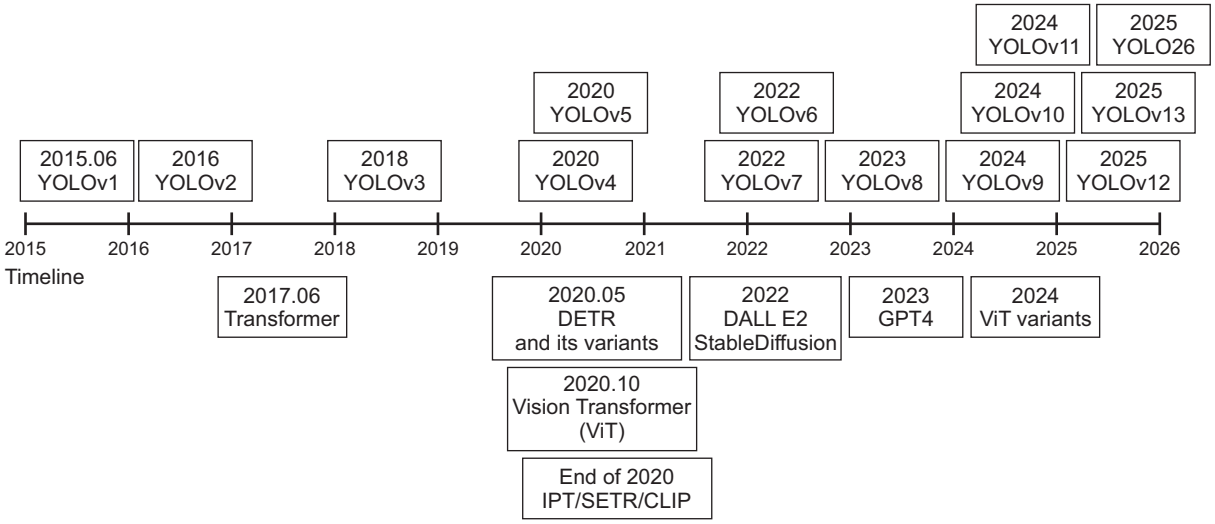

Fig. 1. Key milestones in the development of Vision Transformer and YOLO.

(from YOLOv2 to YOLOv13) have been further optimised for specific needs in the food and agriculture sector. For instance, YOLOv8 simplifies the model structure to accommodate deployment on lightweight edge devices [3], while YOLOv13 enhances multi-scale feature extraction to improve the detection of small objects, such as tiny impurities in seaweed [10].

In contrast, models like ViT [11], Detection Transformer (DETR) [12], Image Processing Transformer (IPT) [13], SegFormer [14], and Swin Transformer [15] have also developed rapidly. Taking the most typical ViT as an example, it replaces the local convolutions of CNNiza with a global self-attention mechanism, becoming a highly disruptive alternative. It is particularly adept at capturing long-range contextual dependencies, such as associating scattered leaf disease spots with the overall severity of crop disease [16]. However, ViT's high computational cost often conflicts with the low-resource, real-time requirements of many agricultural scenarios, such as on-site detection by drones. As the core task supporting the aforementioned applications, object detection has experienced exponential growth, facilitated by decreasing hardware costs. However, in food and agriculture scenarios, the presence of unstructured environments, such as varying field lighting conditions or occlusion of livestock in barns, makes the performance trade-off between ViT and YOLO a critical issue that urgently needs resolution [17].

As shown in Fig. 1, in recent years, object detection algorithms have seen rapid development in both research and application within food and agriculture. Despite this rapid progress, related literature is often fragmented and lacks scientific rigor. This paper aims to fill this gap by comparing the progress and applications of YOLO and ViT in this field, providing a detailed analysis of their advantages and limitations, exploring how they can be integrated with advanced algorithms, and systematically assessing their potential and challenges in object detection.

Methodology

The review process involves comprehensive literature retrieval, collection, careful screening, systematic evaluation, and analysis of the applications of YOLO, ViT, and their variants in the field of food and agriculture. We rigorously follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard to depict the detailed workflow and steps. This includes the literature search procedure, bibliometric review, and systematic literature review, reflecting a logical progression from a "broad initial search" to a "progressively stringent screening" process. This ensures that the final selected literature is both relevant and aligned with the research theme.

As shown in Fig. 2, we selected two authoritative electronic databases, the scientific literature indexing Web of Science (Clarivate, Philadelphia, Pennsylvania, USA) and the search engine Google Scholar (Google, Mountain View, California, USA), and used the search strategy "YOLO" OR "Vision Transformer" AND "agriculture" OR "food" to define the core research focus and application domains of the technology. This initial screening identified 425 documents. Subsequently,

we excluded non-English publications, preprints, and thematically irrelevant literature, narrowing the selection down to 257 documents that better met the criteria. Finally, through a rigorous screening process, we excluded documents that were not peer-reviewed, thematically unrelated to the application of "YOLO/Transformer technology in food and agriculture," unavailable in full text, or studies that employed only a single baseline model. This resulted in 48 documents emerging as the final eligible studies for this systematic review. These selected documents simultaneously satisfy the two key requirements of high thematic relevance and reliable academic quality.

Models in food and agriculture

The ViT architecture, introduced by VASWANI [16], revolutionised the field of sequence modelling by addressing the limitations of recurrent neural networks and long short-term memory. Unlike its predecessors, the ViT relies entirely on self-attention mechanisms to capture dependencies within the input data, enabling efficient parallel processing and eliminating the need for recurrence. This approach significantly enhances training efficiency and model performance across various natural language processing tasks [18]. As illustrated in Fig. 3, the ViT model architecture comprises three core components: multi-head attention, positional encoding, and feed-forward network. Among these, multi-head attention and feed-forward networks are the primary elements that constitute both the encoder and the decoder. Since the ViT does not employ recurrent structures, it requires a method to incorporate positional information into the sequence. Positional encoding is used to provide positional information for each position in the input sequence, which is typically generated via sine and cosine functions.

From an architectural perspective, there are significant differences between ViT and the YOLO architecture commonly used in the food and agriculture domain, as illustrated in Fig. 4. These differences directly determine their suitability for different scenarios: ViT, with its core global self-attention mechanism, can capture the relationships among all elements in the input at once (for example, the spatial dependencies between scattered disease spots and the entire plant in a crop image). In contrast, YOLO relies on the local convolution operations of CNN, extracting local features through sliding windows. While this approach efficiently captures detailed information, it has limited capability in modelling long-range dependencies.

In terms of positional information processing, ViT requires the injection of spatial or temporal information through additional positional encodings. YOLO, on the other hand, leverages the CNN's mechanisms of "progressive expansion of convolutional receptive fields" and "grid-based localisation" to inherently preserve positional features, eliminating the need for a separate positional encoding module. This characteristic also

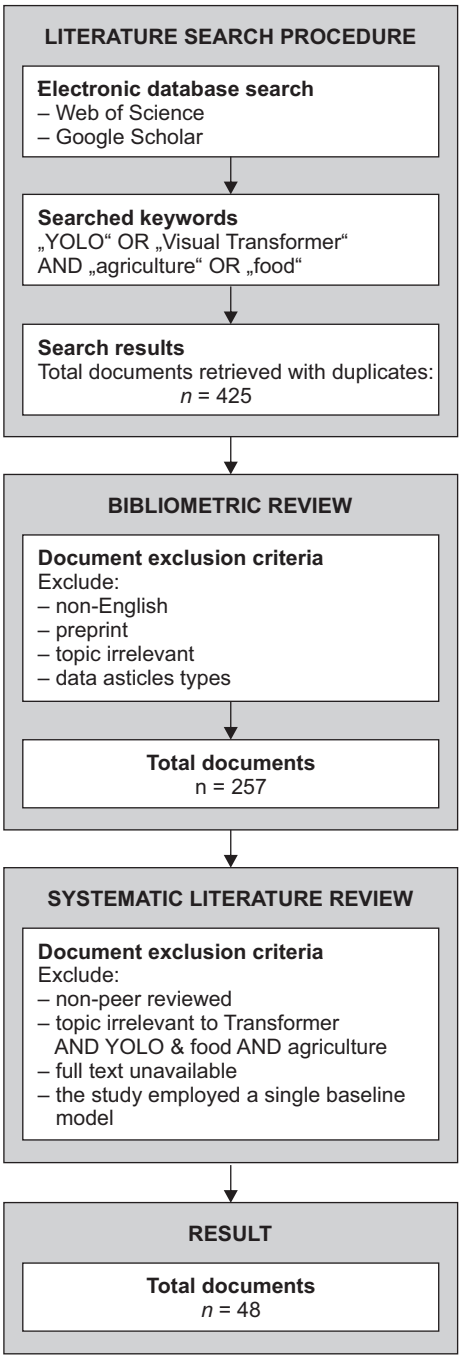

Fig. 2. The methodology of this review.

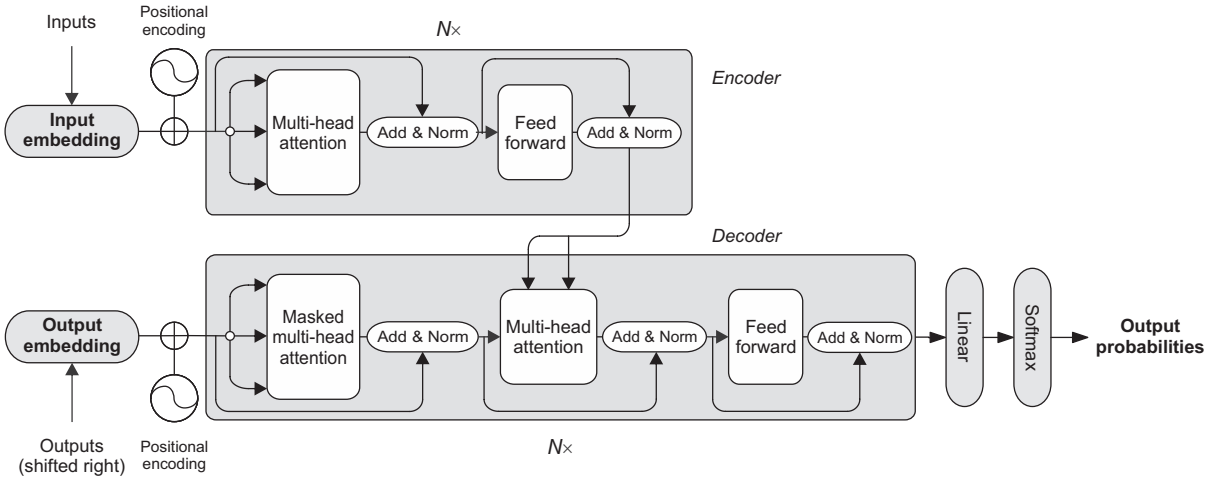

Fig. 3. Architecture of the Transformer model.

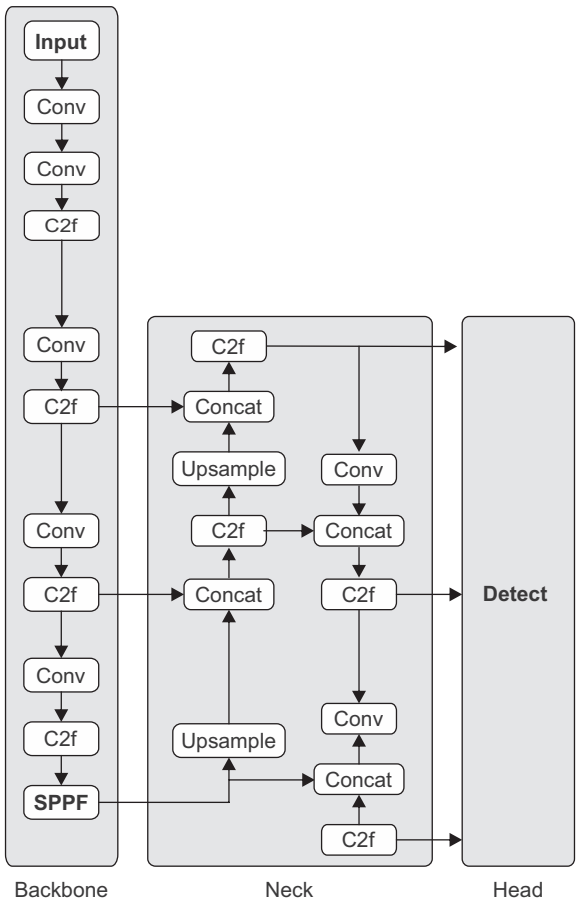

Fig. 4. Architecture of the YOLO model.

Conv – convolution (core of convolutional neural network, extracts visual features, e.g. fruit contour/texture, for YOLO-based detection models; C2f – C2 fusion module (YOLOv8-derived module, enhances multi-scale target expression via parallel feature reuse, lightweight and accurate; SPPF – spatial pyramid pooling - fast (optimised spatial pyramid pooling module, fuses multi-receptive-field features for robust fruit detection in complex orchard scenes).

contributes to its simpler architecture and faster inference speed. Regarding parallelism and real-time performance, the Transformer's self-attention mechanism theoretically supports full parallel computation, but its computational complexity increases quadratically with the input size. YOLO, as a "single-stage detector", achieves higher inference efficiency through its end-to-end regression task design and lightweight convolution kernels, making it more suitable for real-time scenarios in the food and agriculture domain.

ViT, YOLO, and their variants have demonstrated significant potential in food and agriculture. As shown in Fig. 5, their applications span comprehensive scenarios such as fruit quality analysis, animal product classification, crop growth monitoring, and livestock surveillance. Compared to traditional manual observation, these models effectively address challenges in accuracy, practicality, cost and real-time monitoring efficiency. This section will provide a comprehensive review of the applications of YOLO and ViT models across different scenarios.

Fruit quality assessment

Effective fruit ripeness assessment is crucial for optimising harvest timing, improving fruit quality, and reducing post-harvest losses [19]. By accurately determining fruit quality, the optimal harvest time can be identified, maximising yield and minimising waste. The YOLO series demonstrates significant advantages in real-time performance due to its anchor-based detection mechanism and lightweight architecture. In 2024, an improved YOLOv8s model, incorporating Fasternet and Ghost modules, achieved 93.8 % mean average precision (mAP) and a 12.7 % increase in frames

per second (FPS) compared to the original model for field detection of strawberry ripeness [20]. Its lightweight nature allows direct deployment on mobile devices, meeting the needs for rapid field inspections in orchards. In contrast, ViT excels in precisely quantifying ripeness levels. In 2023, a pre-trained ViT-based model for tomato ripeness classification achieved 98.7 % testing accuracy in distinguishing three ripeness stages (unripe – partially ripe – ripe), significantly outperforming CNN models like EfficientNet [21]. Furthermore, it required only one-third of the training samples needed by traditional methods to achieve stable performance.

In practical scenarios, challenges such as complex lighting and occlusions impose dual demands on detection technologies. The instance segmentation versions of the YOLO series show strong adaptability. In 2023, Yue et al. [22] optimised feature fusion in YOLOv8-seg, achieving 45 FPS inference speed and a segmentation Intersection over Union (IoU) of 0.85 on a complex dataset containing both healthy and diseased tomatoes. This enables rapid outlining of overlapping fruit edges, suitable for large-scale field inspections [22]. Meanwhile, the AS-SwinT model, based on the Transformer architecture, utilises a Swin Transformer backbone to extract features from grape berries. Combined with adaptive feature fusion and anchor scale optimisation, it achieved 62.8 % AP_mask in segmenting densely occluded grape clusters, particularly excelling at distinguishing small berries under shadow. Its performance significantly surpassed traditional models like Mask R-CNN [23].

Animal product classification

Accurate detection and classification of animal product quality are fundamental to ensuring food safety and enhancing production efficiency. Traditional chemical analysis and manual inspection suffer from low efficiency and significant time delays. In this field, YOLO and ViT form a distinct technological complementarity: the former excels in real-time performance, while the latter specialises in high-precision analysis, collectively driving the intelligent advancement of animal product detection.

In the classification of major categories such as meat and poultry, YOLO models enable rapid screening at production line levels due to their lightweight architecture. In 2023, an image processing model based on YOLOv8 was proposed for the freshness grading of milkfish (bangus) [24]. By employing multi-stage feature extraction for efficient classification, it provided an accurate and fast detection tool for aquatic product processing lines, with the method being directly adaptable to batch sorting needs in practical production [24]. Another study focusing on aquatic organism detection proposed a lightweight Mobile-YOLO model [25]. Through its Mobile-Nano backbone network and lightweight detection head design, it achieved a 32.2 % parameter reduction while maintaining a 95.2 % FPS and a 1.6 % mAP improvement over the baseline model in detecting four types of aquatic targets, including sea cucumbers and sea urchins. This fully validated the high efficiency of the YOLO series in complex underwater environments [25].

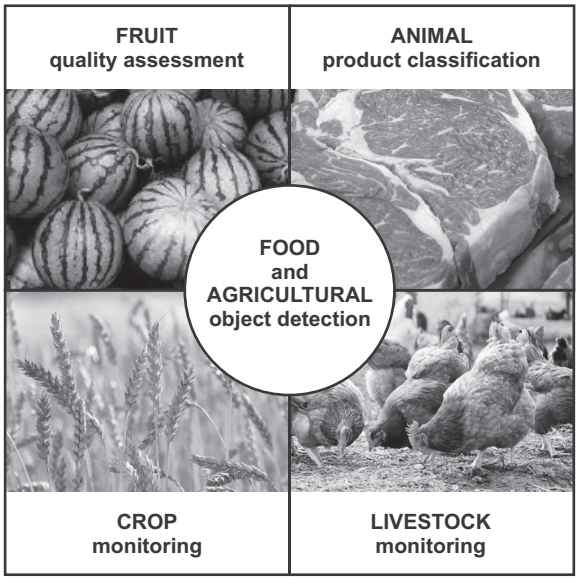

Fig. 5. Practical application scenarios of YOLO and Vision Transformer.

ViT demonstrates significant advantages in capturing fine-grained features. In a study on detecting turning-over behaviour of caged meat ducks, a YOLOv8s-Swin Transformer model was constructed by integrating a Swin Transformer-tiny module into the YOLOv8 backbone. This model achieved a 97.1 % average recognition accuracy on a dataset of 1 000 samples, with a false detection rate of only 2.0 %. Its self-attention mechanism significantly outperformed traditional YOLO models in extracting features of small targets in dense cage environments [26]. In a sheep face detection task, a model named DT-YOLOv5-S, proposed by Guo et al. [27], enhanced global information capture capability via Swin Transformer. It achieved 87.4 % mAP and 61 FPS inference speed on a dataset featuring occlusions and multiple angles, with overall performance surpassing traditional algorithms like

Faster R-CNN and YOLOv4. These two model types exhibit a scenario differentiation: "YOLO is suited for rapid sorting, while ViT is suited for precise feature identification".

Crop monitoring

To enhance crop yield and land use efficiency, modern agriculture widely adopts mixed-cropping systems. However, the varying growth cycles of different crops make traditional classification and monitoring methods difficult to adapt [28]. In this context, the YOLO series, with its advantages in real-time performance and lightweight design, is suitable for rapid field inspections, while ViT breaks through the limitations of complex scenarios with its multi-modal feature fusion and global information capture capabilities, promoting intelligent mixed-cropping monitoring.

In mixed-cropping systems, where crop categories are interwoven and morphologically similar, high demands are placed on balancing the speed and accuracy of classification technologies. The YOLO series achieves real-time field classification with its single-stage detection architecture: a 2023 study proposed a transfer learning-based YOLOv8n model [29]. Optimised via transfer learning from the COCO128 dataset, it achieved a 92.7 % mAP in complex field environments and can be directly transferred to wheat-rapeseed mixed-cropping scenarios, providing a technical reference for rapid classification in similar intercropping systems.

A 2024 study proposed a lightweight CNN-Transformer hybrid network named MCT Net [30]. It integrates Sentinel-2 time-series multispectral data, using a CNN sub-module to extract local spectral features and a Transformer submodule to capture temporal dependencies. It also designs an attention-guided learnable positional encoding module to address issues of data missing caused by cloud cover. In experiments conducted in Arkansas wheat-barley mixed-cropping areas, this model achieved an overall accuracy of 96.8 % and a Kappa coefficient of 0.951, representing a 12.3 % improvement over ResNet-50. It proved particularly adept at classifying gramineous crops with overlapping growth periods. In 2022, a study proposed a Multi-branch Self-learning Vision Transformer (Msvit), which fused SAR and optical time-series data to construct a multi-modal feature space [31]. In complex intercropping scenarios, its feature utilisation rate was 21 % higher than YOLOv8, with the classification error controlled within 4.7 %, significantly outperforming single-modal models. These two model types exhibit a scenario differentiation; YOLO is suited for rapid field statistics, while ViT is suited for precise multi-modal classification.

Livestock monitoring

Commercial free-range farming systems also face challenges such as complex backgrounds, multi-scale targets, occlusion, and posture variations [32]. In livestock farming, behavioural monitoring is essential for preventing disease transmission and achieving contactless health assessment, yet traditional manual monitoring is limited by low efficiency and high costs. The emergence of computer vision has effectively addressed these issues.

Taking multi-scale and occlusion challenges in flock detection as an example, the YOLO series enables real-time monitoring in farms through its lightweight architecture: In 2023, a study proposed an improved YOLOv8-based solution incorporating squeeze-and-excitation attention mechanisms and Mosaic data augmentation [33]. This model achieved 94.7 % mAP at IoU = 0.5 on the public Chicken-Gender Dataset with an inference speed of 52 FPS, making it suitable for inspection robots in farming environments.

ViT, on the other hand, enhances robustness in occluded scenarios through feature interaction design. In 2024, an enhanced ViT baseline model named EMSC-DETR, built upon the RT-DETR framework with optimised feature interaction, achieved 94.3 % mAP on the public PoultryDet-2024 dataset-representing a 2.8 percentage point improvement over YOLOv8 [34].

For clear and intuitive reading and understanding, this review provides Tab. 1, a comparative benchmark table for YOLO and ViT baselines, summarising model parameter comparisons and their respective performance across various datasets and scenarios.

Challenges

Despite the significant potential demonstrated by YOLO and ViT baselines and their variants in food and agricultural object detection, they simultaneously face substantial challenges.

Common challenges

Agricultural and food scenarios suffer from high sample annotation costs and imbalanced class distribution, such as the "long-tail distribution" of agricultural data and multi-environment variations. Factors like different crop growth stages and changes in food appearance due to lighting and morphological variations lead to insufficient data diversity. This can cause decreased model robustness and limited generalisation ca-

Tab. 1. YOLO and Vision Transformer comparative benchmark.

	Usage scenario		
	Object detection	Realtime segmentation	Image classification
Advantage			
Vision Transformer	Strong fine-grained classification capability, able to distinguish similar diseases.	Good robustness against complex backgrounds.	High potential in few-shot learning.
YOLO	Suitable for real-time computing in the field and deployable on mobile devices.	Capable of simultaneous multi-object detection, ideal for fruit counting and pest detection.	Low hardware requirements, making it suitable for agricultural field environments.
Dataset	COCO2017	UEC-FOODPIX	VOC-2017
Params(M) [10^6]			
Vision Transformer	12.2	55.9	41.9
YOLO	8.9	54.6	36.5
GFLOPs(G) [10^9]			
Vision Transformer baseline	12.3 (DETR)	8.6 (Twins-B)	129.6 (RT-DETR)
YOLO baseline	7.0 (YOLOv4-S)	None	103.2 (YOLOv7)
Reference	[45]	[46]	[47]

Params(M) – number of model parameters, GFLOPs(G) – model computational complexity (expressed as the number of floating-point operations).

pability due to differences in food processing environments, posing challenges for any algorithm [35].

Prevalent agricultural automation equipment and food sorting pipelines require detection models to possess low-latency characteristics. The hard requirement for "millisecond-level response" in agricultural scenarios forces models to optimise computational efficiency. Striking a balance between accuracy and speed is a common challenge for different types of models [36].

Distinct challenges

The self-attention mechanism leads to computational complexity that grows quadratically with the input sequence length, resulting in low efficiency in large-scale agricultural scenes (e.g., high-resolution aerial images of farmland). On identical datasets, ViT inference time can be over three times longer than YOLOv5 when processing images with resolutions above 800 × 800 [37]. Reliance on large-scale training data to realise the global modelling advantages of self-attention leads to weak generalisation capability in small-sample agricultural scenarios (e.g., detection of niche crops) [38].

The convolutional-based architecture offers relatively manageable computational complexity, making it easier to meet real-time requirements. However, it tends to lose fine-grained details during large-scale feature fusion, and the multi-scale feature pyramid still has limitations in capturing detailed features of small objects. While its anchor box design and multi-scale training strategy make it easier to optimise with small sample data, it is generally less sensitive to data volume than ViT [39]. YOLO's feature extraction is biased towards locality, leading to insufficient expressive power in fine-grained classification tasks (e.g., food ingredient recognition, crop variety distinction). The YOLO series faces a more severe challenge with class confusion phenomena compared to ViT in fine-grained crop variety classification tasks.

Future prospects

Common future prospects

Self-supervised learning combined with domain adaptation techniques can be employed. This strategy utilises unlabelled agricultural and food data for model pre-training, followed by fine-tuning with limited labelled data to adapt to specific scenarios. Adversarial learning can facilitate cross-domain generalisation for food detection models, enabling adaptation from controlled laboratory settings to real-world complex environments [40].

Developing lightweight Transformer variants, such as using Linear Attention or Sparse Attention mechanisms, or designing CNN-Transformer

Tab. 2. Visual Transformer models and their improvements in object detection.

Variant types	Improvements	Description	Ref.
DETR (Detection Transformer)	End-to-end training	DETR proposes an end-to-end framework that eliminates the need for hand-crafted components in traditional object detection methods.	[12]
	Hungarian algorithm for matching	Use the Hungarian algorithm to match predicted boxes with ground truth boxes, enabling a more robust assignment mechanism.	
	Global context modelling	The attention mechanism allows for effective modelling of global context, improving detection in complex scenes.	
	Feature fusion across domains	By leveraging the transformer architecture, DETR integrates features from different levels, enhancing the detection of small and occluded objects.	
ViT (Vision Transformer)	Patch-based input	Split the input image into fixed-size patches and linearly embed them, allowing the model to process images similarly to sequences in natural language processing tasks.	[11]
	Position encoding	To retain spatial information, introduce learnable position embeddings for the patches.	
	Training strategy	ViT demonstrates strong performance when pre-trained on large datasets, benefiting from larger models and extended training durations.	
	Global feature learning	The self-attention mechanism enables ViT to capture global dependencies effectively, which enhances its performance on various vision tasks.	
IPT (Image Processing Transformer)	Focus on low-level computer Vision tasks	IPT concentrates on low-level computer vision tasks, such as denoising, super-resolution, and deraining.	[13]
	Leveraging large-scale datasets	By leveraging the ImageNet benchmark to generate a large number of corrupted image pairs, the capabilities of the transformer are maximised.	
	Multi-head and multi-tail structure	The IPT model incorporates a multi-head and multi-tail design to enhance its feature learning ability.	
	Introduction of contrastive learning	The introduction of contrastive learning enables the model to better adapt to different image processing tasks.	
Segformer	Simple and efficient framework	SegFormer integrates Transformer with lightweight multi-layer perceptron decoders to create a simple yet powerful semantic segmentation framework.	[14]
	Novel hierarchical encoder	The framework features a hierarchically structured Transformer encoder that outputs multiscale features and does not require positional encoding, thus avoiding performance degradation.	
	Simplified decoder	SegFormer eliminates complex decoder designs by utilising an multi-layer perceptron decoder that aggregates information from different layers, combining both local and global attention to enhance feature representation.	
	Efficiency and performance improvement	The model scales from SegFormer-B0 to SegFormer-B5, achieving significantly better performance and efficiency compared to previous methods.	
	Performance improvement	For instance, SegFormer-B4 achieves 50.3 % mean Intersection over Union (mIoU) on ADE20K (with 64 million parameters), making it 5 times smaller and 2.2 % better than the previous best method. The best model, SegFormer-B5, reaches 84.0 % mIoU on the Cityscapes validation set and demonstrates excellent zero-shot robustness on Cityscapes-C.	
TNT (Transformer in Transformer)	Fine-grained feature extraction	TNT further divides the input image's local patches into smaller patches ("visual words"), allowing the model to better capture features of objects at different scales and locations.	[48]
	Local attention mechanism	TNT emphasises the calculation of attention within local patches, enhancing the model's performance in handling complex images.	
	Low-cost attention calculation	TNT can compute attention between each visual word and other words in the visual sentence at a low computational cost, improving efficiency.	
	Enhanced representation ability	By aggregating the features of both visual words and visual sentences, TNT improves the overall representation capability of the model.	
	Outstanding performance	Experimental results show that TNT achieves 81.5 % top-1 accuracy on ImageNet and outperforms state-of-the-art visual Transformer with similar computational costs.	

Tab. 2. *continued*

Variant types	Improvements	Description	Ref.
Swin Transformer	Hierarchical architecture	It employs a hierarchical structure that enables feature modelling at various scales, enhancing its adaptability to different visual tasks.	[15]
	Shifted window mechanism	The implementation of shifted windows allows self-attention computation to be confined to non-overlapping local windows, improving computational efficiency while facilitating cross-window connections.	
	Linear computational complexity	The computational complexity of Swin Transformer is linear with respect to image size, making it more efficient for processing high-resolution images compared to traditional Transformer, which can become computationally expensive with larger images.	
	Broad applicability	Swin Transformer demonstrates compatibility with a wide range of vision tasks, including image classification, object detection, and semantic segmentation, achieving state-of-the-art performance across these tasks.	
	Performance improvement	It significantly outperforms traditional Transformer, achieving higher accuracy metrics, particularly in object detection and semantic segmentation tasks, thus highlighting its effectiveness as a visual backbone.	

ADE20K Dataset – a widely used benchmark dataset (ADE – architecture, detection, and embedding) for scene semantic segmentation jointly released by institutions including Massachusetts Institute of Technology (Cambridge, Massachusetts, USA) and Stanford University (Stanford, California, USA). It is designed to evaluate the performance of models in understanding complex real-world scenes.

hybrid architectures can help balance global modelling capabilities with computational efficiency. Integrating the local feature extraction strengths of CNN with the global contextual understanding of ViT can achieve an optimal balance of "accuracy and speed" in target sorting applications [41].

Tab. 2 lists visual Transformer models in object detection and their respective advancements for readers' reference.

Future directions for distinct challenges

Techniques like sparse attention mechanisms (e.g., localised window attention) or knowledge distillation can be adopted to reduce computational complexity. For instance, knowledge distillation could allow lightweight YOLO models to incorporate the robust global feature representation capabilities characteristic of ViT [42].

Anchor-free designs or dynamic feature fusion mechanisms can be implemented to enhance the capture of fine-grained details. Eliminating predefined anchor boxes and incorporating pixel-level classification may improve localisation accuracy. Furthermore, attention-guided dynamic feature fusion could potentially increase the recall rate for detecting small targets in agricultural and food contexts [43].

Enhancing model robustness in complex scenes can be achieved by integrating complementary tasks like semantic segmentation or leveraging multi-spectral data. For example, adding a segmentation head to the YOLO architecture to create a multi-task "detection and segmentation" framework could significantly increase its practical utility [44].

CONCLUSIONS

This paper provides a detailed comparison of YOLO and ViT algorithms, outlining key milestones in the evolution of their models. It also introduces the architectures of YOLO and ViT, analysing their respective strengths and weaknesses in object detection within food and agriculture. Furthermore, the performance and efficiency of these two types of algorithms across diverse and complex application scenarios are examined. Experimental results indicate that the YOLO series offers fast detection speed, meets real-time processing demands, and consumes relatively fewer resources, making it suitable for deployment on edge devices. However, it suffers from relatively lower accuracy and limited ability to distinguish fine features and complex backgrounds. In contrast, the ViT series achieves high detection accuracy in complex scenes, is less affected by target occlusion, and exhibits strong scalability. Nonetheless, these advantages come with drawbacks such as high computational complexity, slow inference speed, and challenges in real-time performance. This study provides a theoretical foundation and practical guidance for algorithm selection and optimisation in real-world applications.

Acknowledgements

This study was supported by the Natural Science Foundation of Fujian Province of China (Grant No. 2022J01821 and No. 2022J05163), the National Key R&D Program of China (Grant No. 2020YFD0900904 and No. 2023YFD2100603), and the National Natural Science Foundation of China (Grant No. 11705068 and No. 32172339).

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.

REFERENCES

1. Wang, B.: Automatic mushroom species classification model for foodborne disease prevention based on Vision Transformer. Journal of Food Quality, *2022*, 2022, article 1173102. ISSN: 1745-4557. DOI: 10.1155/2022/1173102.
2. Abbaspour-Gilandeh, Y. – Aghabara, A. – Davari, M. – Mája, J. M.: Feasibility of using computer vision and artificial intelligence techniques in detection of some apple pests and diseases. Applied Sciences, *12*, 2022, article 906. ISSN: 2076-3417. DOI: 10.3390/app12020906.
3. Gao, Z. – Huang, J. – Chen, J. – Shao, T. – Ni, H. – Cai, H.: Deep transfer learning-based computer vision for real-time harvest period classification and impurity detection of *Porphyra haitnensis*. Aquaculture International, *32*, 2024, pp. 5171–5198. ISSN: 0967-6120. DOI: 10.1007/s10499-024-01422-6.
4. Gao, Z. – Chen, S. – Huang, J. – Cai, H.: Real-time quantitative detection of hydrocolloid adulteration in meat based on Swin Transformer and smartphone. Journal of Food Science, *89*, 2024, pp. 4359–4371. ISSN: 0022-1147. DOI: 10.1111/1750-3841.17159.
5. Ren, S. – He, K. – Girshick, R. – Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, *39*, 2017, pp. 1137–1149. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2577031.
6. Lin, T.-Y. – Goyal, P. – Girshick, R. – He, K. – Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. New York : Institute of Electrical and Electronics Engineers, 2017, pp. 2999–3007. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.324.
7. He, K. – Gkioxari, G. – Dollár, P. – Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. New York : Institute of Electrical and Electronics Engineers, 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.
8. Ning, C. – Zhou, H. – Song, Y. – Tang, J.: Inception single shot multibox detector for object detection. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China. New York : Institute of Electrical and Electronics Engineers, 2017, pp. 549–554. ISBN: 978-1-5386-0560-8. DOI: 10.1109/ICMEW.2017.8026312.
9. Redmon, J. – Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA. New York : Institute of Electrical and Electronics Engineers, 2017, pp. 6517–6525. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.690.
10. Zhang, S. – Guo, X. – Tan, M. – Yang, C. – Wang, Z. – Li, G. – Wang, B.: DE-YOLOv13-S: Research on a biomimetic Vision-based model for yield detection of yunnan large-leaf tea trees. Biomimetics, *10*, 2025, article 724. ISSN: 2313-7673. DOI: 10.3390/biomimetics10110724.
11. Dosovitskiy, A. – Beyer, L. – Kolesnikov, A. – Weissenborn, D. – Zhai, X. – Unterthiner, T. – Dehghani, M. – Minderer, M. – Heigold, G. – Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 2021 Proceedings of the International Conference on Learning Representations (ICLR 2021), Vienna, Austria. OpenReview.net [online], published 12 January 2021. <https://openreview.net/pdf?id=YicbFdNTTy>
12. Carion, N. – Massa, F. – Synnaeve, G. – Usunier, N. – Kirillov, A. – Zagoruyko, S.: End-to-end object detection with Transformers. In: Vedaldi, A. – Bischof, H. – Brox, T. – Frahm, J. M. (Eds): Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12346. Cham : Springer, 2020, pp. 213–229. ISBN: 978-3-030-58451-1. DOI: 10.1007/978-3-030-58452-8_13.
13. Chen, H. – Wang, Y. – Guo, T. – Xu, C. – Deng, Y. – Liu, Z. – Ma, S. – Xu, C. – Xu, C. – Gao, W.: Pre-trained image processing transformer. In: 2021 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Nashville, USA. New York : Institute of Electrical and Electronics Engineers, 2021, pp. 12294–12305. ISBN: 978-1-6654-4509-2. DOI: 10.1109/CVPR46437.2021.01212.
14. Xie, E. – Wang, W. – Yu, Z. – Anandkumar, A. – Alvarez, J. M. – Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Ranzato, M. – Beygelzimer, A. – Dauphin, Y. – Liang, P. S. – Wortman Vaughan, J. (Eds.): Advances in neural information processing systems 34 (NeurIPS 2021). Red Hook : Curran Associates, 2021, pp. 12077–12090. ISBN: 9781713845393. <https://papers.nips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf>
15. Liu, Z. – Lin, Y. – Cao, Y. – Hu, H. – Wei, Y. – Zhang, Z. – Lin, S. – Guo, B.: Swin transformer: Hierarchical Vision Transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada. New York : Institute of Electrical and Electronics Engineers, 2021, pp. 9992–10002.

ISBN: 978-1-6654-2812-5. DOI: 10.1109/ICCV48922.2021.00986.
16. Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A. N. – Kaiser, Ł. – Polosukhin, I.: Attention is all you need. In: Guyon, I. – Von Luxburg, U. – Bengio, S. – Wallach, H. – Fergus, R. – Vishwanathan, S. – Garnett R. (Eds.): Advances in neural information processing systems 30 (NIPS 2017). Red Hook : Curran Associates, 2017, pp. 6000–6010. ISBN: 9781510860964. <https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
17. Ariza-Sentís, M. – Vélez, S. – Martínez-Peña, R. – Baja, H. – Valente, J.: Object detection and tracking in precision farming: a systematic review. Computers and Electronics in Agriculture, *219*, 2024, article 108757. ISSN: 0168-1699. DOI: 10.1016/j.compag.2024.108757.
18. Devlin, J. – Chang, M.-W. – Lee, J. – Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis : Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
19. Li, B. – Lecourt, J. – Bishop, G.: Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction – a review. Plants, *7*, 2018, article 3. ISSN: 2223-7747. DOI: 10.3390/plants7010003.
20. Liu, R. – Wang, Q. – Zhang, H. – Wang, L.: An efficient and lightweight YOLOv8s strawberry maturity detection model. In: ASIG '24: Proceedings of the 2024 2nd Asia Symposium on Image and Graphics. New York : Association for Computing Machinery, 2024, pp. 14–22. DOI: 10.1145/3718441.3718444.
21. Nahak, P. – Pansuriya, K. – Pratihar, D. K. – Deb, A. K.: Vision transformer-based transfer learning approach for tomato maturity stage classification. In: Proceedings of the 15th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2023). Lecture Notes in Networks and Systems, vol. 1245. Cham : Springer, 2023, pp. 118–127. ISBN: 978-3-031-81082-4. DOI: 10.1007/978-3-031-81083-1_11.
22. Yue, X. – Qi, K. – Na, X. – Zhang, Y. – Liu, Y. – Liu, C.: Improved YOLOv8-Seg network for instance segmentation of healthy and diseased tomato plants in the growth stage. Agriculture, *13*, 2023, article 1643. ISSN: 2077-0472. DOI: 10.3390/agriculture13081643.
23. Du, W. – Liu, P.: Instance segmentation and berry counting of table grape before thinning based on AS-SwinT. Plant Phenomics, *5*, 2023, article 0085. ISSN: 2643-6515. DOI: 10.34133/plantphenomics.0085.
24. Samaniego, L. A. – Peruda, S. R. – Brucal, S. G. E. – Yong, E. D. – De Jesus, L. C. M.: Image processing model for classification of stages of freshness of bangus using YOLOv8 algorithm. In: 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan. New York : Institute of Electrical and Electronics Engineers, 2023, pp. 401–403. ISBN: 979-8-3503-4019-8. DOI: 10.1109/GCCE59613.2023.10315381.
25. Jiang, H. – Zhao, J. – Ma, F. – Yang, Y. – Yi, R.: Mobile-YOLO: A lightweight object detection algorithm for four categories of aquatic organisms. Fishes, *10*, 2025, article 348. ISSN: 2410-3888. DOI: 10.3390/fishes10070348.
26. Md Akbar, J. U. – Kamarulzaman, S. F.: YOLOv8s-Swin: Enhanced tomato ripeness detection for smart agriculture. International Journal of Advanced Computer Science and Applications, *16*, 2025, pp. 1006–1014. ISSN: 2158-107X. DOI: 10.14569/IJACSA.2025.0160897.
27. Guo, Y. – Yu, Z. – Hou, Z. – Zhang, W. – Qi, G.: Sheep face image dataset and DT-YOLOv5s for sheep breed recognition. Computers and Electronics in Agriculture, *211*, 2023, article 108027. ISSN: 0168-1699. DOI: 10.1016/j.compag.2023.108027.
28. Wu, B. – Zhang, M. – Zeng, H. – Tian, F. – Potgieter, A. B. – Qin, X. – Yan, N. – Chang, S. – Zhao, Y. – Dong, Q. – Boken, V. – Plotnikov, D. – Guo, H. – Wu, F. – Zhao, H. – Deronde, B. – Tits, L. – Loupian, E.: Challenges and opportunities in remote sensing-based crop monitoring: a review. National Science Review, *10*, 2023, article nwac290. ISSN: 2095-5138. DOI: 10.1093/nsr/nwac290.
29. Vini, S. L. – Rathika, P.: Automated tomato leaf disease identification via transfer learning approach on convolutional neural network. In: 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Krishnankoil, India. New York : Institute of Electrical and Electronics Engineers, 2024, pp. 01–05. ISBN: 979-8-3503-6119-3. DOI: 10.1109/INCOS59338.2024.10527708.
30. Wang, Y. – Feng, L. – Sun, W. – Wang, L. – Yang, G. – Chen, B.: A lightweight CNN-Transformer network for pixel-based crop mapping using time-series Sentinel-2 imagery. Computers and Electronics in Agriculture, *226*, 2024, article 109370. ISSN: 0168-1699. DOI: 10.1016/j.compag.2024.109370.
31. Li, K. – Zhao, W. – Peng, R. – Ye, T.: Multi-branch self-learning Vision Transformer (MSViT) for crop type mapping with Optical-SAR time-series. Computers and Electronics in Agriculture, *203*, 2022, article 107497. ISSN: 0168-1699. DOI: 10.1016/j.compag.2022.107497.
32. Okinda, C. – Nyalala, I. – Korohou, T. – Okinda, C. – Wang, J. – Achieng, T. – Wamalwa, P. – Mang, T. – Shen, M.: A review on computer vision systems in monitoring of poultry: a welfare perspective. Artificial Intelligence in Agriculture, *4*, 2020, pp. 184–208. ISSN: 2589-7217. DOI: 10.1016/j.aiia.2020.09.002.
33. Wu, D. – Ying, Y. – Zhou, M. – Pan, J. – Cui, D.: Improved ResNet-50 deep learning algorithm for identifying chicken gender. Computers and Electronics in Agriculture, *205*, 2023, article 107622. ISSN: 0168-1699. DOI: 10.1016/j.compag.2023.107622.

34. Li, X. – Cai, M. – Tan, X. – Yin, C. – Chen, W. – Liu, Z. – Wen, J. – Han, Y.: An efficient transformer network for detecting multi-scale chicken in complex free-range farming environments via improved RT-DETR. Computers and Electronics in Agriculture, *224*, 2024, article 109160. ISSN: 0168-1699. DOI: 10.1016/j.compag.2024.109160.
35. Khan, Z. – Shen, Y. – Liu, H.: Object detection in agriculture: A comprehensive review of methods, applications, challenges, and future directions. Agriculture, *15*, 2025, article 1351. ISSN: 2077-0472. DOI: 10.3390/agriculture15131351.
36. Kim, J. – Kim, G. – Yoshitoshi, R. – Tokuda, K.: Real-time object detection for edge computing-based agricultural automation: a case study comparing the YOLOX and YOLOv12 architectures and their performance in potato harvesting systems. Sensors, *25*, 2025, article 4586. ISSN: 1424-8220. DOI: 10.3390/s25154586.
37. Parez, S. – Dilshad, N. – Alghamdi, N. S. – Alanazi, T. M. – Lee, J. W.: Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. Sensors, *23*, 2023, article 6949. ISSN: 1424-8220. DOI: 10.3390/s23156949.
38. Hamidisepehr, A. – Mirnezami, S. V. – Ward, J. K.: Comparison of object detection methods for corn damage assessment using deep learning. Transactions of the ASABE, *63*, 2020, pp. 1969–1980. ISSN: 2151-0040. DOI: 10.13031/trans.13791.
39. Wang, C.-Y. – Bochkovskiy, A. – Liao, H.-Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Vancouver, Canada. New York : Institute of Electrical and Electronics Engineers, 2023, pp. 7464–7475. ISBN: 979-8-3503-0130-4. DOI: 10.1109/CVPR52729.2023.00721.
40. Luo, Y. – Liu, W. – Li, H. – Lu, Y. – Lu, B.-L.: A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. Journal of Neural Engineering, *21*, 2024, article 046004. ISSN: 1741-2552. DOI: 10.1088/1741-2552/ad546d.
41. Padshetty, S. – Umashetty, A.: Agricultural innovation through deep learning: a hybrid CNN-Transformer architecture for crop disease classification. Journal of Spatial Science, 2024, pp. 1–32. ISSN: 1449-8596. DOI: 10.1080/14498596.2024.2355225.
42. Han, Y. – Huang, Z. – Sun, Y. – Wang, B. – Chen, Q.: Agricultural object detection in complex environments via co-attention and self-knowledge distillation. Information Sciences, *724*, 2025, article 122711. ISSN: 0020-0255. DOI: 10.1016/j.ins.2025.122711.
43. Yan, H. – Guo, H. – Wei, L. – Xu, X. – Liang, Y. – Li, Y. – Chen, S. – Yu, P.: A global feature fusion and adaptive optimization method to enhance detection accuracy and computational efficiency based on YOLOv8. Alexandria Engineering Journal, *129*, 2025, pp. 538–552. ISSN: 1110-0168. DOI: 10.1016/j.aej.2025.06.025.
44. Yang, T. – Zhou, S. – Xu, A. – Ye, J. – Yin, J.: YOLO-SegNet: a method for individual street tree segmentation based on the improved YOLOv8 and the SegFormer network. Agriculture, *14*, 2024, article 1620. ISSN: 2077-0472. DOI: 10.3390/agriculture14091620.
45. Fang, Y. – Liao, B. – Wang, X. – Fang, J. – Qi, J. – Wu, R. – Niu, J. – Liu, W.: You Only Look at one sequence: rethinking transformer in vision through object detection. In: Ranzato, M. – Beygelzimer, A. – Dauphin, Y. – Liang, P. S. – Wortman Vaughan, J. (Eds.): Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Red Hook : Curran Associates, 2021, pp. 26183–26197. ISBN: 9781713845393. <https://papers.nips.cc/paper_files/paper/2021/file/dc912a253d1e9ba40e2c597ed2376640-Paper.pdf>
46. Okamoto, K. – Yanai, K.: UEC-FoodPIX Complete: A large-scale food image segmentation dataset. In: Del Bimbo, A. – Cucchiara, R. – Sclaroff, S. – Farinella, G. M. – Mei, T. – Bertini, M. – Escalante, H. J. – Vezzani, R. (Eds.): Pattern Recognition. ICPR International Workshops and Challenges. Lecture Notes in Computer Science, vol. 12665. Cham : Springer, 2021, pp. 647–659. ISBN: 978-3-030-68821-9. DOI: 10.1007/978-3-030-68821-9_51.
47. Lu, J. – Zhao, Y. – Yu, M.: PGLD-YOLO: a lightweight algorithm for pomegranate fruit localisation and recognition. PeerJ Computer Science, *11*, 2025, article e3307. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.3307.
48. Han, K. – Xiao, A. – Wu, E. – Guo, J. – Xu, C. – Wang, Y.: Transformer in Transformer. In: Ranzato, M. – Beygelzimer, A. – Dauphin, Y. – Liang, P. S. – Wortman Vaughan, J. (Eds.): Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Red Hook : Curran Associates, 2021, pp. 15908–15919. <https://papers.nips.cc/paper_files/paper/2021/file/854d9fca60b4bd07f9bb215d59ef5561-Paper.pdf>

Received 30 June 2025; 1st revised 4 November 2025; accepted 11 December 2025; published online 18 December 2025.