Pertanika J. Sci. & Technol. 34 (3): 1331 - 1354 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

YOLOv9-ResCBAM: Enhancing Tomato Ripeness Detection in Greenhouses Through Advanced Object Detection

Jalal Uddin Md Akbar[1], Syafiq Fauzi Kamarulzaman[1]*, Muhammad Danial Mohamad Rizwan[1], Riadul Islam Rabbi[2], and Ekramul Haque Tusher[1]

[1]Faculty of Computing, Universiti Malaysia Pahang, Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

ABSTRACT

Accurate object detection and classification are pivotal in precision agriculture for tasks such as identifying crop varieties and assessing ripeness stages. To optimise the yield and quality of tomatoes within the variable conditions of a greenhouse, it is crucial to accurately detect and classify their ripeness levels. This study introduces YOLOv9-ResCBAM, an enhanced object detection model based on the advanced YOLOv9 architecture, designed to classify tomato ripeness levels (fully ripe, partially ripe, and unripe) in greenhouse environments. Firstly, a comprehensive tomato image dataset reflecting diverse greenhouse conditions was curated. Secondly, rigorous preprocessing techniques, including auto-orientation, resizing, and augmentation, were applied to enhance dataset quality. Finally, the proposed YOLOv9-ResCBAM model was trained and evaluated. The findings indicated that the proposed model achieved a superior mean Average Precision (mAP@0.5) of 0.912, outperforming both earlier YOLO-based detectors and other established object detection frameworks like SSD, Mask R-CNN, and Faster R-CNN. This improvement is attributed to innovations like Programmable Gradient Information (PGI), Generalised Efficient Layer Aggregation Network (GELAN), and our integration of a Residual Convolutional Block Attention Module (ResCBAM). YOLOv9-ResCBAM's exceptional performance in accurately detecting and categorising tomatoes across ripeness stages, even in challenging greenhouse scenarios, provides a promising advancement for agricultural object detection.

ARTICLE INFO

Article history:
Received: 22 June 2025
Accepted: 27 March 2026
Published: 12 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.03

E-mail addresses:
jalaluddinmdakbar00@gmail.com (Jalal Uddin Md Akbar)
syafiq29@umpsa.edu.my (Syafiq Fauzi Kamarulzaman)
danialrizwan1108@gmail.com (Muhammad Danial Mohamad Rizwan)
riadul.rabbi72@gmail.com (Riadul Islam Rabbi)
ekramulhaquetusher@gmail.com (Ekramul Haque Tusher)
* Corresponding author

e-ISSN: 2231-8526

This study serves as a foundational step toward AI-driven solutions in crop management, enabling more efficient resource allocation and ultimately contributing to more sustainable and optimised food production practices.

Keywords: Agricultural automation, attention mechanism, computer vision, object detection, smart agriculture, YOLO, YOLOv9

INTRODUCTION

Advancements in Computer Vision (CV) algorithms, particularly in object detection, have revolutionised numerous industries, including agriculture. The application of Convolutional Neural Networks (CNNs) marked the beginning of significant strides in this domain (Akbar et al., 2024). Among the notable breakthroughs in this field, the You Only Look Once (YOLO) series stands out for its revolutionary approach to detecting objects in a single forward pass, markedly enhancing processing speeds and detection accuracy. Throughout its evolution, each iteration of the YOLO architecture has been developed with the goal of advancing the model's performance in terms of both computational efficiency and accuracy. YOLOv9 (Wang et al., 2024), developed by Chien-Yao Wang and colleagues, is the latest in this innovative lineage, offering substantial improvements over its predecessors. This version introduces two novel concepts: Programmable Gradient Information (PGI) and Generalised Efficient Layer Aggregation Network (GELAN). Building upon these advancements, we propose YOLOv9-ResCBAM, an enhanced model that integrates a Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 architecture to further improve detection accuracy.

Crop ripeness detection and classification have emerged as crucial endeavours in the realm of precision agriculture, enabling informed decision-making and optimised yield management (Rizzo et al., 2023). The conventional approach for these tasks has been manual inspection, a method that is not only time-consuming and labour-intensive but is also subject to inconsistencies and human error. The advent of computer vision and deep learning techniques has introduced opportunities for automation, offering accurate, real-time, and scalable solutions to address these challenges.

Early approaches to crop ripeness detection relied on methods like colour and texture analysis (Akbar et al., 2023; Moreira et al., 2022; Wang et al., 2023; Yang et al., 2023). Despite their potential, the reliability of these methods was frequently compromised by environmental variables, including fluctuating lighting conditions, shadows, and occlusions. These factors consequently constrained their applicability and robustness for large-scale agricultural operations.

With the rapid advancements in deep learning and computer vision techniques, researchers have increasingly turned to data-driven approaches for crop ripeness

detection and classification. Convolutional Neural Networks (CNNs) function as powerful instruments for identifying and extracting hierarchical visual features from images, a capability that facilitates accurate object detection and classification tasks. In the context of crop ripeness detection, approaches that utilise deep learning have proven to be more effective than conventional methods (Phoophuangpairoj et al., 2023; Tang et al., 2023). These techniques can learn intricate patterns and representations directly from labelled data, allowing them to adapt to the complexities of real-world agricultural environments. However, the effectiveness of deep learning models depends greatly on the accessibility of high-quality, diverse, and annotated datasets, which can be challenging and resource-intensive to obtain.

In this study, Tomatoes were chosen due to their economic importance and the complexities involved in accurately detecting their ripeness stages. As one of the most widely grown vegetables globally, tomatoes are a critical component of the agricultural production and food supply chains (Brodt et al., 2013). Their ripeness directly affects the market value, taste, and nutritional content, making accurate detection crucial for both farmers and consumers (Cui et al., 2022). Additionally, tomatoes present a variety of ripeness indicators, including colour, texture, and firmness, which pose challenges to detection systems. These complexities make tomatoes an ideal subject for evaluating the capabilities of advanced CV algorithms like YOLOv9. By focusing on tomatoes, this study aimed to address a significant agricultural challenge and demonstrate the practical applications of cutting-edge technology in enhancing crop management and production efficiency.

Several investigations have examined the application of YOLO-based models for detecting crop ripeness, yielding promising results. For instance, Li et al. (2023) introduced a tomato maturity detection system utilising YOLOv5, with a high precision and mean Average Precision (mAP) of 0.96 for tomato ripeness classification in greenhouse settings. However, their model might struggle under outdoor conditions owing to variations in lighting and environmental factors. Greenhouses provide controlled environments that are less variable than those of outdoor fields. Hence, the robustness of their model under varying outdoor conditions remains a concern. Our approach aims to enhance adaptability by incorporating additional data augmentation techniques and a more diverse dataset that includes various environmental conditions. Another study by Li et al. (2023) proposed MHSA-YOLOv8, a YOLOv8- based model enhanced with a multi-head self-attention mechanism, for the grading and counting of tomato ripeness in complex environments. Although the model was effective in complex environments, its performance in severe occlusion situations was limited. This limitation is critical, as agricultural fields often present significant occlusions due to foliage.

The model achieved a mean Average Precision (mAP) of 0.94; however, its robustness under occluded conditions remains a challenge. Zeng et al. (2023) came up with a lightweight approach for tomato detection based on enhanced YOLOv5 and MobileNetV3, emphasising real-time detection and mobile implementation for classifying tomato ripeness. The model achieved a mAP of 0.97 and a detection speed of 42.5 ms on CPU platforms. Although the model demonstrated impressive real-time performance, it struggled with dense small-target detection, which is common in crops such as tomatoes, where fruits are often clustered. Su et al. (2022) introduced SE-YOLOv3-MobileNetV1, a lightweight network for tomato maturity classification under natural greenhouse conditions, which achieved a high mAP of 0.92. However, their study did not discuss the performance of their model in real-time detection scenarios. Real-time applications are crucial for practical agricultural applications.

Wang et al. (2023) presented an improved YOLOv5n model with a coordinate attention mechanism and a novel loss function for real-time cherry tomato maturity detection, achieving a mAP of 0.89. However, their research was limited to a specific cherry tomato dataset, raising concerns about the generalisability of the model to other tomato varieties. Our study addresses this limitation by training the model on a broader dataset that includes multiple tomato varieties to ensure wider applicability. Badeka et al. (2023) explored a deep learning approach using YOLOv7 to assess grape maturity in precision viticulture and achieved a mAP of 0.88. However, a limitation of that study was its sole concern with a particular type of grape, and it did not explore how the model would perform with other grape varieties. Chen et al. (2024) developed MTD-YOLOv7, a multi-task deep convolutional neural network based on YOLOv7, for the simultaneous detection of cherry tomato fruit bunches, fruit maturity, and cluster maturity. The model obtained a mean Average Precision (mAP) of 0.90 for maturity detection but faced challenges with densely packed tomatoes. Additionally, to accurately identify blueberry fruits at various ripeness levels, Wang et al. (2024) developed the YOLO-BLBE model. Their approach involved optimising the YOLOv5s model by adding a colour enhancement algorithm, which resulted in a mean Average Precision (mAP) of 0.91. Despite these advancements, future work is needed to identify semi-mature blueberries at different growth stages.

Recent research has explored the integration of attention mechanisms into object detection models to enhance their ability to focus on relevant features. For instance, Chien et al. (2024) combined object detection with attention mechanisms to detect paediatric wrist fractures, demonstrating improved performance. In a related study, Zamri et al. (2024) improved the detection of small drones by optimising a YOLOv8 model with the integration of attention mechanisms. However, the potential of combining such mechanisms with the latest YOLO architecture for agricultural applications, particularly in complex greenhouse environments, remains largely unexplored.

Recent studies have also successfully explored attention mechanisms for dense tomato detection; for instance, Appe et al. (2023a) utilised CAM-YOLO, while Appe et al. (2023b) integrated coordinate attention into YOLO architectures to improve dense fruit classification. However, standard attention modules can sometimes suffer from gradient degradation during deep feature extraction. Our integration of ResCBAM specifically addresses this by combining spatial and channel attention with a residual connection. This design works synergistically with YOLOv9's Programmable Gradient Information (PGI) to preserve baseline feature fidelity while actively mitigating the gradient vanishing issues seen in earlier attention-based YOLO variants, making it uniquely suited for heavily occluded greenhouse environments.

These studies collectively demonstrate the efficacy of YOLO-based models for crop ripeness detection, revealing areas for improvement, particularly in greenhouse environments with occlusions, variable lighting, and dense plant distributions. While attention mechanisms have shown promise in computer vision, their potential in YOLO-based models for agriculture remains underexplored. This motivates our proposal of YOLOv9-ResCBAM, which integrates attention mechanisms into YOLOv9 to address these challenges.

Applying YOLOv9-ResCBAM in greenhouses, especially for tomato detection, presents unique challenges due to controlled but varied conditions. This study explores how YOLOv9, enhanced by attention mechanisms, improves detection accuracy and adaptability in diverse agricultural scenarios. It aims to demonstrate the potential of cutting-edge computer vision technologies to enhance crop management and production in precision agriculture. The primary contributions of this study are:

1. A novel enhancement to the YOLOv9 architecture by integrating a Residual Convolutional Block Attention Module (ResCBAM), improving detection accuracy in complex agricultural environments.
2. Improved classification of tomato ripeness across varying stages, addressing challenges like occlusions, dense foliage, and lighting variations commonly found in greenhouses.
3. Demonstrated performance through rigorous evaluation, achieving an mAP of 0.912, outperforming earlier YOLO versions and traditional object detection models like SSD, Faster R-CNN, and Mask R-CNN.
4. Established an advancement in agricultural object detection, with potential for widespread application across different crops and environments.

MATERIALS AND METHODS

This study introduces our proposed YOLOv9- ResCBAM model, which integrates a Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 architecture, for tomato ripeness detection in greenhouses. Our workflow, illustrated in Figure 1,

begins with dataset preparation, crucial for training and evaluating all models. The dataset undergoes preprocessing and augmentation stages, including label extraction, annotation conversion to YOLO format, auto-orientation, resizing, class label simplification, and filtering. We employed Roboflow for data augmentation, applying techniques such as random flips, rotations, shears, blurring, and noise addition to simulate diverse greenhouse conditions.

The processed data was split into training (92%), validation (4%), and test (4%) sets, ensuring comprehensive model evaluation. The proposed YOLOv9-ResCBAM and other object detection models were trained on this prepared dataset and assessed for their effectiveness in detecting tomatoes at various ripeness stages under diverse agricultural conditions.

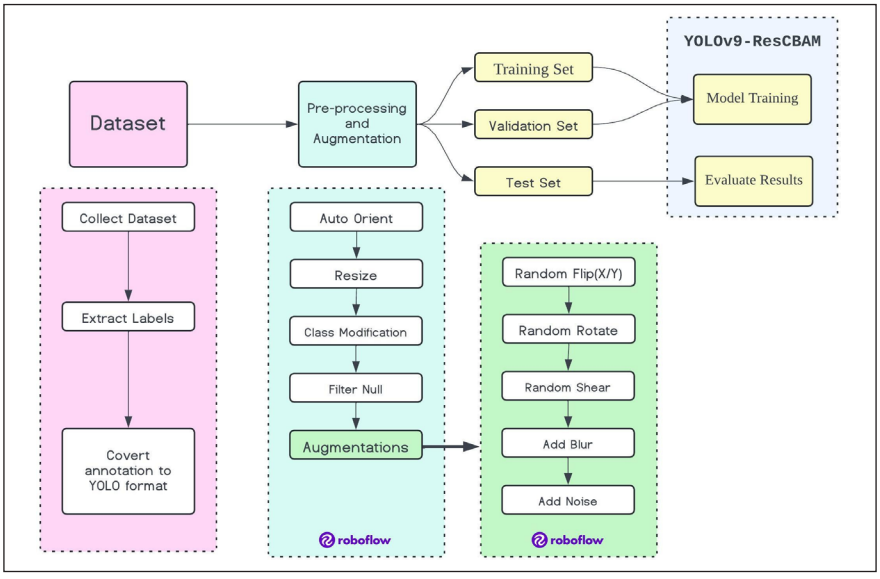


Figure 1. Workflow diagram for YOLOv9-ResCBAM model training

Proposed Method: YOLOv9-ResCBAM

YOLOv9-ResCBAM is an advanced object detection architecture that builds upon the YOLOv9 framework by integrating the Residual Convolutional Block Attention Module (ResCBAM). Figure 2 depicts the proposed architecture. This integration enhances the model's ability to focus on critical parts of the input image, thereby improving detection accuracy. The primary innovation in this model is the incorporation of ResCBAM, which combines channel and spatial attention mechanisms with residual learning to refine feature maps at multiple stages within the network.

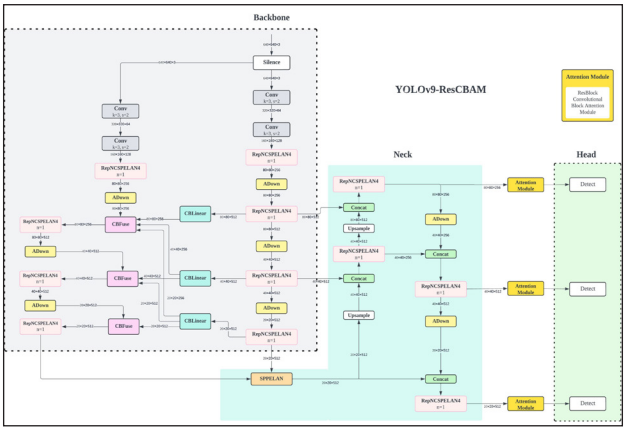


Figure 2. YOLOv9-ResCBAM architecture diagram

Attention Module

In deep learning models, attention mechanisms play an essential role, particularly in object detection, by enabling the model to concentrate on the most significant parts of the input. The use of these mechanisms has recently achieved notable results in the object detection field (Guo et al., 2022; Yao et al., 2022; Zamri et al., 2024). YOLOv9-ResCBAM employs the Convolutional Block Attention Module (CBAM), which improves the model's performance by prioritising significant features and lowering the influence of less important ones. Figure 3 demonstrates that CBAM comprises two sequential sub-modules: the Channel Attention Module (C-Attention) and the Spatial Attention Module (S-Attention).

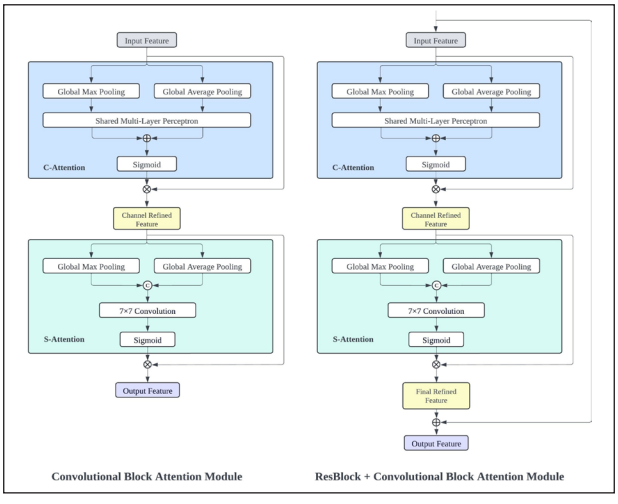


Figure 3. Detailed representation of the Convolutional Block Attention Module (CBAM). The left side depicts the standard CBAM architecture, while the right side illustrates the modified architecture with a residual connection, known as ResCBAM

Channel Attention (C-Attention)

The primary function of the Channel Attention mechanism is to focus on the inter-channel relationships of the feature maps. It allows the network to prioritise the channels that contain the most informative features. Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width of the feature map, respectively, the channel attention mechanism utilises two global pooling operations: Global Average Pooling (GAP) and Global Max Pooling (GMP). GAP aggregates spatial information into a channel descriptor by averaging each channel's spatial dimension, while GMP captures the most salient features across the spatial dimensions by selecting the maximum value within each channel. These pooled descriptors are then passed through a shared Multi-Layer Perceptron (MLP), which consists of a single hidden layer to generate the channel attention map M_C, formally defined in Equation 1:

$$M_C(F) = \sigma(MLP(GAP(F)) + MLP(GMP(F))) \quad [1]$$

Here, σ is the sigmoid activation function. This attention map is then applied to the original feature map via element-wise multiplication to generate the channel-refined feature map F_C according to Equation 2:

$$F_C = M_C(F) \otimes F \quad [2]$$

Where $\otimes$ denotes the element-wise multiplication. This operation ensures that the channels with higher attention values contribute more significantly to the next layers.

Spatial Attention (S-Attention)

Channel attention and spatial attention serve distinct but complementary roles; while channel attention is concerned with determining 'what' features are significant, spatial attention identifies 'where' those important features are located on the feature map. The Spatial Attention Module in CBAM refines the feature map by emphasising crucial regions within the feature map. Given the channel-refined feature map F_C, the spatial attention is computed as follows:

The spatial attention mechanism first applies Global Average Pooling (GAP) and Global Max Pooling (GMP) across the channel axis, resulting in two 2D maps, F_C^{avg} and F_C^{max}, both of size $H \times W$.

Subsequently, the two maps are combined through channel-wise concatenation and then fed into a convolutional layer that features a 7 × 7 kernel. This operation produces the spatial attention map M_S, as expressed in Equation 3:

$$M_S(F_C) = \sigma\left(f^{7\times7}([GAP(F_C); GMP(F_C)])\right) \quad [3]$$

Here, $f^{7\times7}$ represents the convolution operation utilising a 7 × 7 filter, while σ denotes sigmoid function that scales the output to the range [0, 1]. This spatial attention map M_S highlights the regions within the feature map that are most informative for the task at hand.

The final spatially refined feature map F_S is then obtained by applying the spatial attention map to the channel-refined feature map, formally defined in Equation 4:

$$F_S = M_S(F_C) \otimes F_C \quad [4]$$

This operation directs the network to prioritise the most salient essential locations within the feature map.

Convolutional Block Attention Module (CBAM)

CBAM combines the Channel and Spatial Attention mechanisms sequentially to refine the feature maps both in the channel and spatial domains. The following equation summarises the complete operation of the CBAM module according to Equation 5:

$$F_{CBAM} = M_S(M_C(F) \otimes F) \otimes (M_C(F) \otimes F) \quad [5]$$

Where F_{CBAM} is the output feature map that has been refined by both channel and spatial attention mechanisms. By applying both types of attention, CBAM effectively filters out irrelevant features while amplifying important ones.

Residual Connection in ResCBAM

To further enhance the performance and stabilise the learning process, a residual connection is incorporated into CBAM, forming the Residual Convolutional Block Attention Module (ResCBAM). The residual connection is formulated as Equation 6:

$$F_{output} = F + F_{CBAM} \quad [6]$$

This equation implies that the input features are added directly to the output of the CBAM module. This residual connection helps in preserving the original information from the feature map while adding the benefits of attention, which prevents overfitting and improves generalisation.

Integration in YOLOv9

We strategically place ResCBAM modules at key points in the YOLOv9 architecture, particularly in the neck of the network, as shown in Figure 2. This specific insertion depth was selected because it represents the junction where multi-scale feature maps are richest in both spatial resolution and semantic depth, allowing the attention mechanism to maximise feature refinement immediately before the detection heads. This placement allows the attention mechanism to refine features after they have been processed by the backbone and before they are used for detection. The ResCBAM module operates on the output of each RepNCSPELAN4 block in the neck, enhancing the feature maps before they are concatenated or used for detection. The enhanced feature map F' after applying ResCBAM is given by Equation 7 :

$$F' = ResCBAM(F) = F + M_s(M_c(F) \otimes F) \otimes (M_c(F) \otimes F) \quad [7]$$

where F is the output feature map from the RepNCSPELAN4 block.

Integrating ResCBAM into YOLOv9 enhances feature refinement, enabling improved ripeness detection through adaptive, multi-scale attention, better gradient flow, and spatial awareness under varied greenhouse conditions and occlusions.

Data Acquisition

The effectiveness of computer vision systems in agricultural settings is heavily dependent on the availability of diverse and high-quality datasets that accurately represent the varied conditions typical of real-world scenarios. For our research, we have selected the Laboro Tomato dataset (https://github.com/laboroai/LaboroTomato) as the data source. This dataset is particularly valuable because it encompasses a comprehensive range of tomatoes at different ripening stages, reflecting the inherent diversity found within a typical agricultural environment. The Laboro Tomato dataset includes images captured using two different cameras, each varying in resolution and image quality. This dual-source approach enriches the dataset with a variety of imaging conditions that closely mimic the variability encountered in greenhouse operations. The dataset categorises the images of tomatoes into three ripening stages. These stages are visually represented in Figure 4 as: fully ripened, half-ripened, and green. The fully ripened tomatoes display over 90% red coloration, half- ripened tomatoes exhibit between 30-89% red coloration, and green tomatoes show 0-30% red coloration, with the percentages providing a rough guide due to natural variability. This stratification is critical for our computer vision model, which needs to accurately identify and classify each stage of tomato maturity to support timely and precise agricultural decision-making.

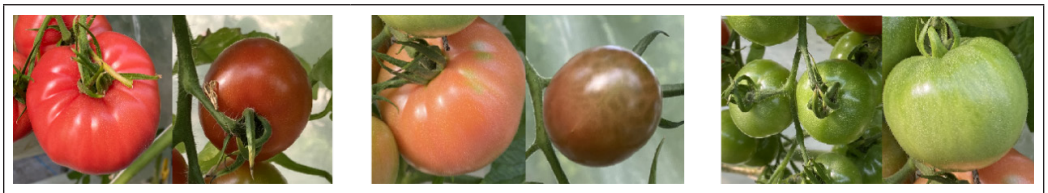


Figure 4. Dataset class

Data Pre-processing and Preparation

Upon acquiring the raw images from the Laboro Tomato dataset, a rigorous preprocessing journey was undertaken to tailor the data for our development pipeline. This process began with a meticulous image selection phase, where images like sample A and sample B, as seen in Figure 5, were reviewed for quality and relevance. Subsequently, a series of rigorous preprocessing steps was applied to the Laboro Tomato dataset to optimise it for our models. Key steps included:

1. Auto Orient : Images were realigned using EXIF orientation metadata to ensure consistency.
2. Resize : All images were uniformly resized to 640 × 640 pixels, balancing detail preservation with computational efficiency.
3. Class Modification : We simplified the classification system by focusing solely on ripeness stages, removing size distinctions between tomatoes. Table 1 details the class remapping. By removing size distinctions, we deliberately forced the model to learn generalised colour-based ripeness cues across different phenotypes, rather than risking the model falsely associating 'large' strictly with 'mature,' which would cause failures on naturally smaller tomato cultivars. Furthermore, the discrete three-class ripeness system was chosen over continuous regression because agricultural robotics and harvest planning require strict, actionable decision boundaries. During annotation, the 30-89% coloration rule served as a strict guideline. To mitigate subjectivity, ambiguous borderline fruits were systematically downgraded to the less-mature category to prevent costly premature harvesting errors.
4. Filter Null : Images lacking annotations were removed to maintain dataset integrity.

5. Dataset Split : The dataset was split into training (92%, 2001 images), validation (4%, 86 images), and test (4%, 87 images) sets. The data split was performed randomly at the image level. While the validation and test partitions represent 4% each, the risk of overfitting or data leakage was severely mitigated by the aggressive augmentation pipeline (e.g., severe shearing, up to 1.5px blur, and noise). This mathematically forced the model to learn invariant biological features rather than memorising spatial clusters or camera-specific sensor artefacts. The success of this regularisation is empirically proven in Figure 10, where validation losses tightly track training losses without divergence.

6. Data Augmentation : We applied various techniques using Roboflow. As depicted in Figure 6, a variety of data augmentation methods were applied to the dataset, including:

- Random flips (X/Y axis)
- Rotations (-15° to $+15^{\circ}$)
- Shear transformations ($\pm 10^{\circ}$)
- Blur effect (up to 1.5 pixels)
- Noise addition (up to 1.56% of pixels)

The augmentation parameters were empirically designed to simulate physical greenhouse challenges. For example, blur (up to 1.5 pixels) and noise (up to 1.56%) mimic lens dust and low-light sensor degradation. Furthermore, while natural spatial relationships exist, clustered foliage is highly chaotic. Utilising synthetic mixup (0.15) and copy-paste (0.3) operations deliberately disrupts predictable background contexts, forcing the network to learn the invariant biological features of the tomatoes rather than relying on spatial memorisation. These preprocessing and augmentation steps ensured a robust, diverse dataset that simulates real-world greenhouse conditions, crucial for training our proposed YOLOv9-ResCBAM model.

Sample A Sample B

Figure 5. Original dataset sample

Table 1
Remapping of original classes to simplified classes in the dataset

Original Class	Override	Included
b_fully_ripened	fully_ripened	Yes
b_green	green	Yes
b_half_ripened	half_ripened	Yes
l_fully_ripened	fully_ripened	Yes
l_green	green	Yes
l_half_ripened	half_ripened	Yes

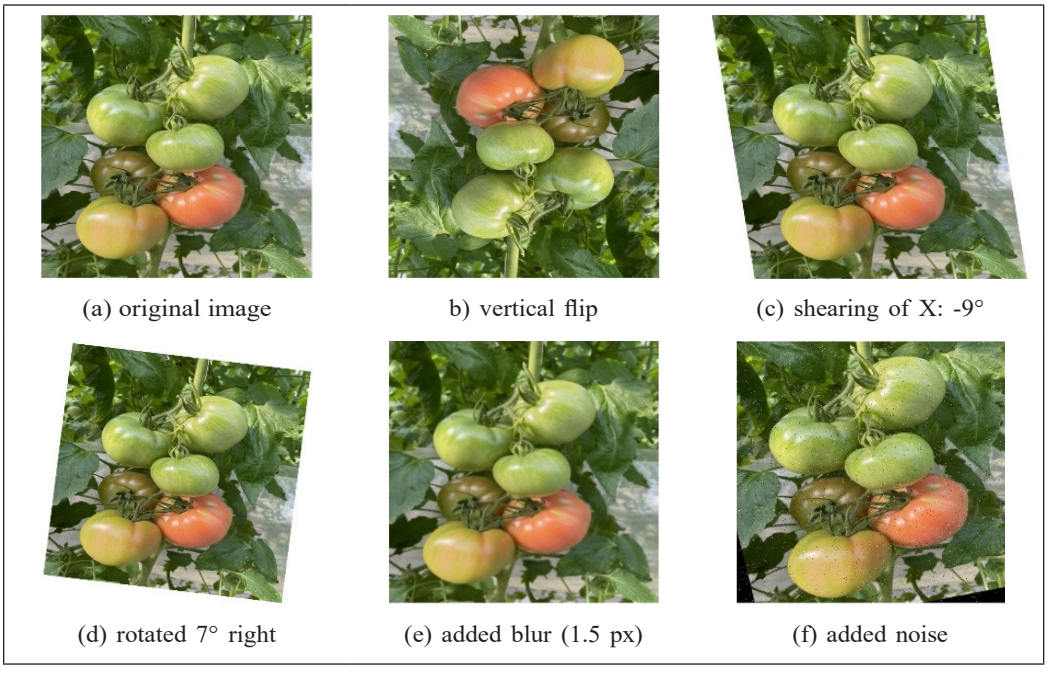
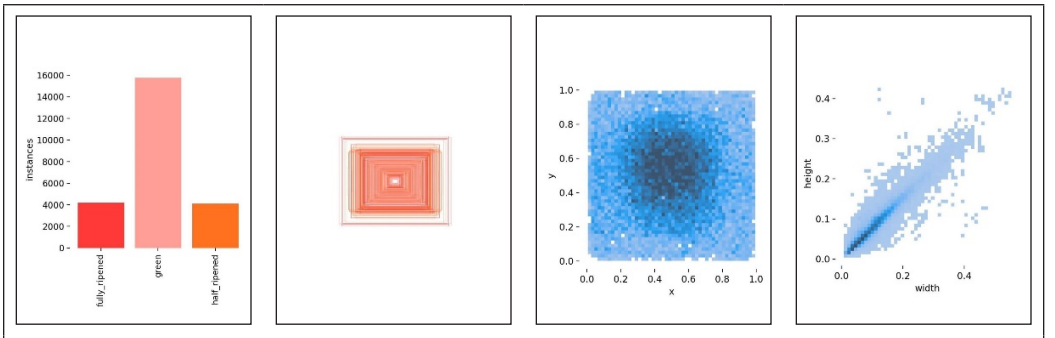


Figure 6. Data augmentation pipeline applied to the dataset; (a) original image; (b) vertical flip; (c) shearing of X: -9°; (d) rotated 7° right; (e) added blur (1.5 px); (f) added noise

Through these meticulous preprocessing steps, the dataset was not only standardised but also enriched, ensuring that it was optimally prepared for training a robust and effective model. Each technique plays a vital role in mirroring real-world variability and challenges, equipping the model with the capability to accurately detect and classify tomatoes across different stages of ripeness under diverse operational conditions in smart greenhouses. Figure 7 illustrates the statistical analysis of the dataset.



a) Number of instances | b) Overall distribution of all annotated bounding boxes | c) Distribution of centre coordinates of all annotated bounding boxes | d) Distribution of size of all annotated bounding boxes

Figure 7. Statistical visualisation of the dataset

Experimental Setup

The experimental hardware configuration included the use of an NVIDIA GeForce RTX 3060 GPU with 16 GB of VRAM, supported by an AMD Ryzen 5 processor with six cores. The software environment was set up on an Ubuntu 24.04 LTS operating system, with all deep learning models trained and evaluated using PyTorch version 2.3.0. The experimental setup was designed to thoroughly evaluate various iterations of the object detection models, and our proposed YOLOv9-ResCBAM. The training was performed with an image size of 640 and a batch size of 8. Specific training parameters were configured to train the YOLO models. Table 2 lists detailed hyperparameters used in the experiments. To ensure absolute fairness and rule out optimisation bias, all comparative baseline models (YOLOv3-v8, SSD, and R-CNN variants) were trained from scratch using the same hardware, hyperparameter configurations (100 epochs, batch size 8), and augmentation pipelines. While a granular empirical ablation study separating the individual effects of YOLOv9, standard CBAM, and the residual connection is slated for future work, the theoretical design of ResCBAM where the residual connection preserves the foundational PGI features is heavily validated by the aggregate performance leap.

Table 2
Detailed training parameters configured for model training

Model Parameter	Value	Parameter Explanation
epoch	100	Total number of epochs for training
lr0	0.01	Initial learning rate, specific to the optimisation method used
lrf	0.01	Final learning rate as a fraction of the initial rate
momentum	0.937	Momentum factor for SGD/1 for Adam
weight decay	0.0005	Weight decay regularisation factor
warmup epochs	3.0	Number of epochs for gradual learning rate increase

Table 2 (*continued*)

Model Parameter	Value	Parameter Explanation
warmup momentum	0.8	Initial momentum during the warm-up phase
warmup bias lr	0.1	Learning rate for bias during warm-up
mixup	0.15	Probability of performing mixup image data augmentation
copy paste	0.3	Probability of using segment copy-paste augmentation

Evaluation Metrics

The evaluation was based on several key metrics, including precision, recall, mean Average Precision (mAP), F1-Score, and frames per second (FPS).

Precision

Precision is defined as the proportion of correctly identified positive samples among all detected positives. It is mathematically expressed as Equation 8:

$$Precision = \frac{TP}{TP+FP} \quad [8]$$

In this context, TP stands for true positives, which are the instances correctly identified as positive, while FP refers to false positive instances that are incorrectly labelled as positive. A higher precision score signifies a more dependable model, as it means the model makes fewer false-positive detections.

Recall

Recall represents the ratio of correctly identified positive samples to the total number of actual positive samples. This metric measures the model's capacity to detect all relevant instances in a dataset. It is mathematically expressed as Equation 9:

$$Recall = \frac{TP}{TP+FN} \quad [9]$$

Here, true positives (TP) refer to the positive instances that were correctly identified by the model, while false negatives (FN) represents the actual positive cases the model overlooked. A high recall value, therefore, suggests that a smaller number of relevant instances were missed, which leads to a more complete detection performance.

mAP (mean Average Precision)

mAP, calculated at an IoU threshold of 0.5, represents the average of the Average Precision (AP) for all classes. It is a standard measure for assessing object detection algorithms, derived from the area under the precision-recall curve.

FPS (Frames Per Second)

This measures the number of images the model can process per second at a batch size of 1, providing an indication of the model's detection speed. These metrics are essential for quantifying the effectiveness and efficiency of each YOLO model iteration in handling real-world object detection tasks.

RESULTS AND DISCUSSION

In this section, the performance outcomes of the YOLOv9-ResCBAM model are presented, focusing on its effectiveness in detecting and classifying tomatoes at various ripeness stages in greenhouse environments.

Performance of YOLOv9-ResCBAM

The YOLOv9-ResCBAM model was evaluated on a dataset of tomatoes in greenhouse environments, which included fully ripened, half-ripened, and unripe tomatoes. The model performance was measured using precision, recall, and mAP at an Intersection over Union (IoU) threshold of 0.5. These metrics are critical for understanding the model's ability to detect and classify tomatoes accurately across different ripeness stages.

The proposed YOLOv9-ResCBAM shows exceptional performance, with an mAP@0.5 of 0.9124, particularly in maintaining high precision and recall across all categories of tomato ripeness. The fully ripened category achieved a precision of 0.936, while green and half-ripened categories attained precisions of 0.904 and 0.897, respectively, as shown in Figure 8. This consistent performance across all ripeness stages underscores the effectiveness of the integrated Residual Convolutional Block Attention Module in enhancing feature extraction and detection accuracy. Furthermore, under stricter localisation criteria, the model demonstrated strong robustness, achieving an mAP@0.5: 0.95 of approximately 0.80, as evidenced by the metric curves in Figure 10.

As indicated by the dataset statistics, Figure 7a, there is a natural class imbalance heavily skewed toward 'green' instances. However, the model maintained exceptional precision across all classes (Fully Ripened: 0.936, Green: 0.904, Half Ripened: 0.897). This demonstrates that the architectural design and augmentation pipeline successfully prevented the model from becoming biased toward the majority class.

The analysis of the precision-recall curve reveals that the YOLOv9- ResCBAM model significantly enhances the detection and classification capabilities of tomatoes in greenhouse settings. Its ability to maintain high precision and recall across varying confidence thresholds and ripeness stages makes it highly reliable for practical applications. The performance of the proposed model is further confirmed by the confusion matrix in Figure 9, which shows high accuracy in classifying tomatoes across various ripeness stages. These results underscore the model's reliability and the effective integration of attention mechanisms. An analysis of the confusion matrix reveals that the model's primary failure mode is confusing 'green' instances with the 'background.' This is biologically expected, as the visual characteristics of unripe green tomatoes share nearly identical colour and texture profiles with the dense surrounding greenhouse foliage and stems. In the context of robotic harvesting, false positives on fully ripened fruit are highly costly, as picking unripe fruit reduces market value. The model's exceptional precision of 0.936 for the fully ripened category directly minimises this risk, providing a highly reliable operational threshold for automated picking systems.

Additionally, Figure 10 presents different metrics, including the training and validation loss curves for bounding box regression, classification, and the overall loss for the proposed YOLOv9-ResCBAM model. The steady decline in these loss values over the training epochs indicates effective learning and convergence of the model. The validation losses closely follow the training losses, suggesting that the models generalise well to unseen data. The YOLOv9-ResCBAM model shows a consistent and sharp decrease in all loss metrics, which aligns with its superior performance observed in the precision-recall analysis. This comprehensive evaluation underscores the robustness and reliability of the YOLOv9-ResCBAM model, making it an excellent choice for precision agricultural tasks.

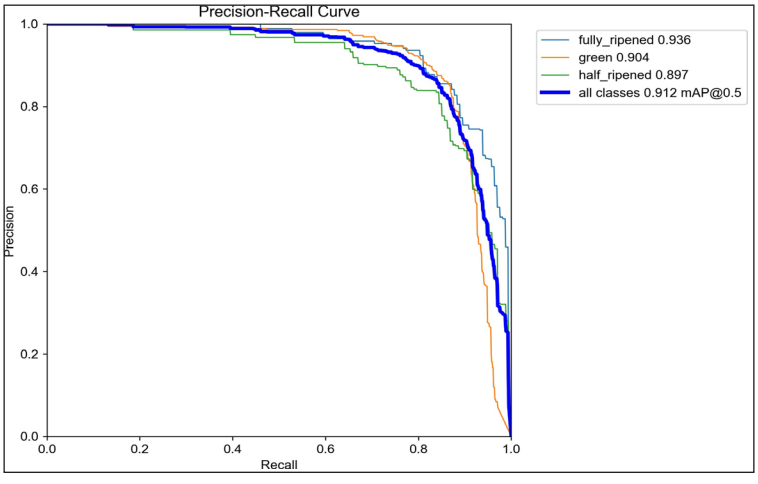


Figure 8. Precision-Recall curve of proposed YOLOv9- ResCBAM model

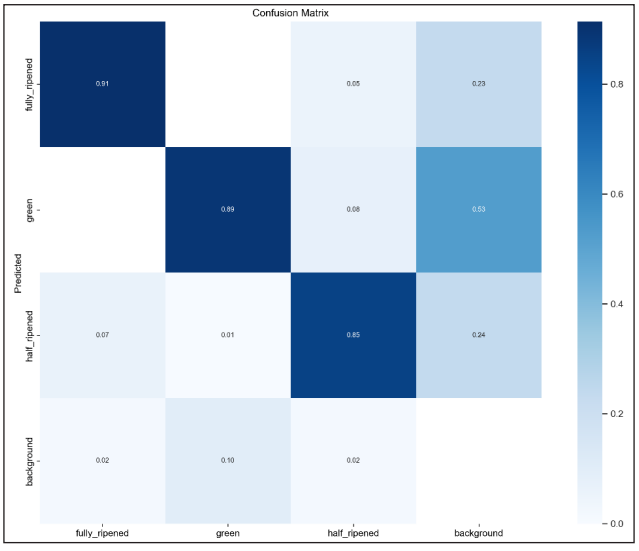


Figure 9. Confusion matrix for YOLOv9-ResCBAM's tomato ripeness classification

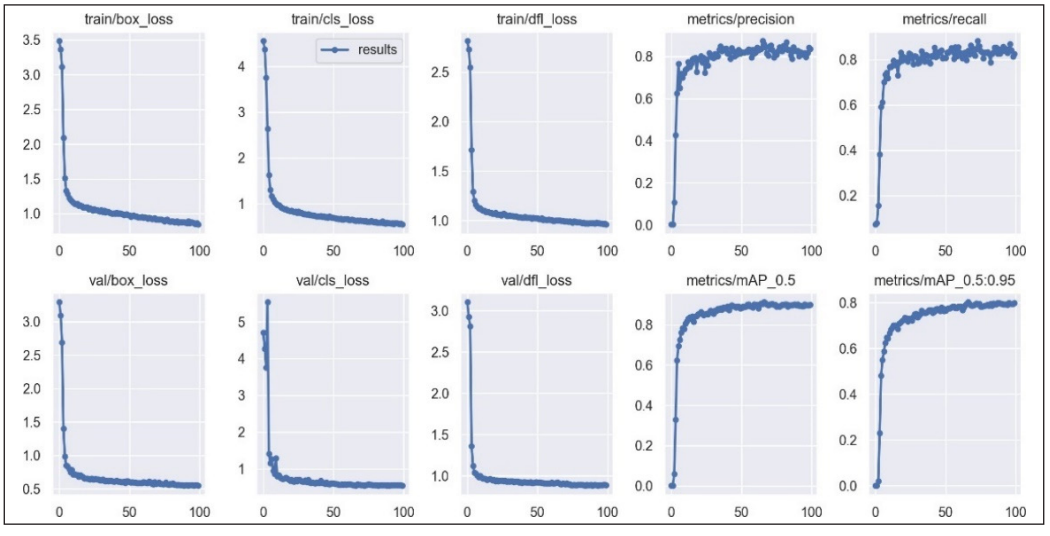


Figure 10. Visualisation of different metrics of the proposed YOLOv9-ResCBAM Model over epochs

DISCUSSION

Accurate detection and classification of crop ripeness stages are pivotal for informed decision-making and optimised yield management in precision agriculture. Through rigorous experimentation, we assessed our proposed YOLOv9-ResCBAM and compared it with previous YOLO iterations and traditional object detection models such as SSD, Faster R-CNN, and Mask R-CNN.

Our findings demonstrate that the proposed YOLOv9-ResCBAM model outperforms all previous models in object detection and classification for agricultural applications. While this study focuses exhaustively on CNN-based, one-stage detectors (such as the YOLO family, SSD, and R-CNN variants), we acknowledge the rapid advancements in Transformer-based detectors (e.g., DETR) for complex computer vision tasks. However, Transformer models currently introduce significant computational overhead and larger memory footprints, which pose substantial challenges for real-time inference on resource-constrained agricultural edge devices. Therefore, CNN-based architectures were prioritised for this benchmarking to ensure practical deployment feasibility, while Transformer-based comparisons remain a highly valuable avenue for future research.

The exceptional performance of YOLOv9-ResCBAM can be attributed to its novel architectural advancements, primarily the integration of the Residual Convolutional Block Attention Module (ResCBAM) into the YOLOv9 framework. This innovation addresses fundamental challenges in deep learning, such as information loss and biased gradient propagation, which have historically impeded optimal performance in complex detection tasks. By incorporating attention mechanisms, ResCBAM ensures that critical information is effectively utilised during the learning process, resulting in more reliable and robust feature representations. This enhancement is particularly beneficial in scenarios where subtle differences between various stages of crop maturity or health need to be accurately recognised and classified. The comparative analysis presented in Table 3 highlights several key insights. YOLOv9-ResCBAM achieved the highest mean average precision (mAP) of 0.912 while maintaining a reasonable Frames Per Second (FPS) rate of 38 on an NVIDIA GeForce RTX 3060 GPU.

As shown in Table 3, while a +0.011 mAP gain over the YOLOv9-c baseline may appear numerically modest, in an automated agricultural context processing thousands of frames, this translates to a tangible reduction in false positives, directly protecting crop yield by preventing the accidental harvesting of unmarketable fruit. Furthermore, while this study focuses on static frame detection, establishing this highly robust frame-level accuracy is the critical prerequisite for future integration with temporal tracking algorithms (such as DeepSORT) across video streams.

This balance between accuracy and speed is crucial for real-time detection applications in greenhouse environments, where timely interventions can significantly impact crop yield and quality. Despite introducing additional computational complexity through its attention mechanisms, YOLOv9-ResCBAM maintains a moderate model size of 117 MB and a GFLOPs (Giga Floating-point Operations per Second) value of 70. This indicates that the model is computationally efficient relative to its high accuracy, making it suitable for deployment on devices with limited processing capabilities.

Table 3
Performance comparison of various models in terms of several key metrics. The FPS values were obtained using an NVIDIA GeForce RTX 3060 GPU

Model	mAP@0.5	Precision	Recall	F1-Score	FPS	Size (MB)	GFLOPs
YOLOv3	0.841	0.869	0.775	0.797	28	118	65
YOLOv4	0.847	0.854	0.802	0.813	25	245	90
YOLOv5	0.892	0.886	0.860	0.845	53	91.4	16.5
YOLOv6	0.876	0.850	0.838	0.828	50	114	18
YOLOv8	0.842	0.834	0.799	0.792	56	83.7	15
YOLO NAS	0.894	0.885	0.864	0.839	20	328	150
YOLOv9-e	0.896	0.876	0.855	0.841	38	115	60
YOLOv9-c	0.901	0.888	0.864	0.847	35	119	60
SSD	0.850	0.851	0.834	0.840	25	103	60
Faster R-CNN	0.880	0.872	0.863	0.841	12	198	120
Mask R-CNN	0.900	0.882	0.875	0.840	10	250	130
YOLOv9-ResCBAM	0.912	0.894	0.882	0.849	38	117	70

Compared to traditional models, YOLOv9-ResCBAM demonstrates superior practicality. SSD, with a mAP of 0.850 and FPS of 25, and Faster R-CNN and Mask R-CNN, with higher mAPs but significantly lower FPS (12 and 10, respectively), are less suited for real-time, resource-constrained environments. The GFLOPs metric further highlights YOLOv9-ResCBAM's optimal balance of computational demand and performance, critical for agricultural settings where hardware resources and energy efficiency are significant considerations. The model's moderate computational requirements facilitate deployment on edge devices common in agriculture, ensuring integration without substantial hardware investments. Although physical latency testing on specific edge hardware, such as the NVIDIA Jetson Nano or Raspberry Pi was outside the immediate scope of this algorithmic study, the theoretical metrics strongly support real-world edge deployment. The proposed model's computational demand of 70 GFLOPs and moderate size of 117 MB fall well within the operational envelopes of modern agricultural edge computing devices, ensuring that high accuracy can be maintained without inducing severe thermal throttling or requiring prohibitive hardware investments. Its adaptability allows customisation for various platforms, enhancing scalability across different agricultural operations. It is important to note that domain shifts across different seasons, variations in farmer pruning styles, and external validation across completely unseen greenhouse datasets remain significant challenges. While our aggressive augmentation pipeline acted as a proxy to simulate environmental shifts, formal cross-dataset validation is a primary focus for our future field trials.

Figure 11 illustrates different challenging situations successfully detected by the YOLOv9-ResCBAM model. The images highlight the model's ability to accurately identify tomatoes under various conditions, such as occlusion by leaves and fruits, dense clustering, and varying lighting conditions. These examples underscore the robustness of the YOLOv9-ResCBAM model in real-world agricultural environments, demonstrating the effectiveness of its integrated attention mechanisms in handling complex scenarios.

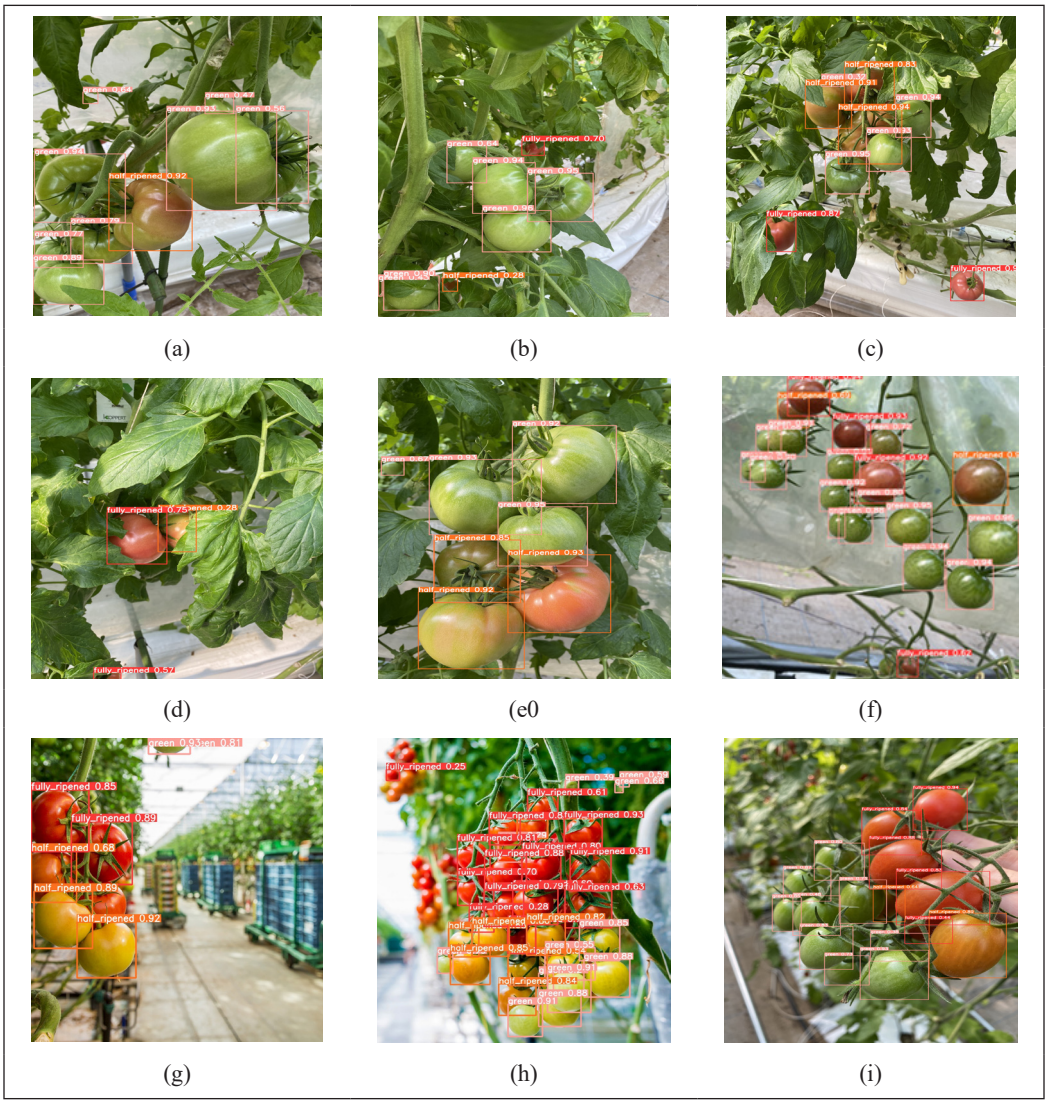


Figure 11. YOLOv9-ResCBAM's robust detection in diverse greenhouse conditions. Examples of tomato fruit detection challenges under different field conditions: (a) fruits occluded by other fruits; (b) fruits occluded by foliage; (c) combined occlusion by foliage and fruits; (d) detection amidst dense foliage; (e) identifying tomato clusters; (f) clustered fruit detection; (g) detection in high-density fruit environments; (h) accurate detection in densely packed fruits; and (i) high-density fruit clustering detection

The seamless integration of cutting-edge computer vision techniques, as exemplified by the YOLOv9-ResCBAM architecture, with domain-specific agricultural knowledge has paved the way for transformative AI-driven solutions in crop management and yield optimisation. Our findings reinforce the potential of the YOLOv9 series, particularly the YOLOv9-ResCBAM model, to revolutionise precision agriculture practices, enabling more efficient resource allocation, timely interventions, and improved productivity and sustainability in greenhouse farming operations.

Furthermore, extending this system to outdoor agricultural settings or multimodal configurations (such as RGB-NIR) presents distinct challenges that warrant future investigation. Outdoor environments introduce highly dynamic illumination, harsh shadows, and complex backgrounds (e.g., sky and soil) that require more aggressive domain adaptation strategies. While integrating Near-Infrared (NIR) imaging could mitigate some lighting variations, it introduces challenges regarding strict multispectral sensor alignment, increased computational load, and higher implementation costs for farmers.

CONCLUSION

This study introduced YOLOv9-ResCBAM, an enhanced object detection model designed for tomato ripeness classification in greenhouse environments. Achieving a mean Average Precision (mAP@0.5) of 0.912, YOLOv9-ResCBAM outperforms earlier YOLO versions and traditional object detection frameworks such as SSD, Faster R-CNN, and Mask R-CNN. Key innovations, including Programmable Gradient Information (PGI), Generalised Efficient Layer Aggregation Network (GELAN), and the Residual Convolutional Block Attention Module (ResCBAM), optimise feature extraction, gradient propagation, and computational efficiency, enabling the model to perform well in challenging greenhouse scenarios. This advancement contributes significantly to agricultural object detection, offering the potential for AI-driven solutions to improve crop management. Future work will involve deploying YOLOv9-ResCBAM in real-world greenhouse scenarios, validating its practical effectiveness, and exploring its applicability to other crops, thereby advancing the broader field of precision agriculture.

ACKNOWLEDGEMENT

We want to extend our heartfelt gratitude to the UMPSA Postgraduate Research Grants Scheme (PGRS) for providing the Open Access funding that supported this research.

REFERENCES

Akbar, J. U. M., Kamarulzaman, S. F., Muzahid, A. J. M., Rahman, M. A., & Uddin, M. (2024). A comprehensive review of deep learning-assisted computer vision techniques for smart greenhouse agriculture. *IEEE Access, 12*, 4485-4522. https://doi.org/10.1109/ACCESS.2024.3349418

Akbar, J. U. M., Kamarulzaman, S. F., & Tusher, E. H. (2023). Plant stem disease detection using machine learning approaches. *2023 14th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-8. https://doi.org/10.1109/ICCCNT56998.2023.10307074

Appe, S. N., Arulselvi, G., & Balamurugan, G. N. (2023a). CAM-YOLO: Tomato detection and classification based on improved YOLOv5 using combining attention mechanism. *PeerJ Computer Science, 9*, Article e1463. https://doi.org/10.7717/peerj-cs.1463

Appe, S. N., Arulselvi, G., & Balamurugan, G. N. (2023b). Detection and classification of dense tomato fruits by integrating coordinate attention mechanism with the YOLO model. In *Advances in computational intelligence and robotics book series* (pp. 278-289). IGI Global. https://doi.org/10.4018/978-1-6684-8098-4.ch016

Badeka, E., Karapatzak, E., Karampatea, A., Bouloumpasi, E., Kalathas, I., Lytridis, C., Tziolas, E., Tsakalidou, V. N., & Kaburlasos, V. G. (2023). A deep learning approach for precision viticulture, assessing grape maturity via YOLOv7. *Sensors, 23*(19), Article 8126. https://doi.org/10.3390/s23198126

Brodt, S., Kramer, K. J., Kendall, A., & Feenstra, G. (2013). Comparing environmental impacts of regional and national-scale food supply chains: A case study of processed tomatoes. *Food Policy, 42*, 106-114. https://doi.org/10.1016/j.foodpol.2013.07.004

Chen, W., Liu, M., Zhao, C., Li, X., & Wang, Y. (2024). MTD-YOLO: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. *Computers and Electronics in Agriculture, 216*, Article 108533. https://doi.org/10.1016/j.compag.2023.108533

Chien, C., Ju, R., Chou, K., Xieerke, E., & Chiang, J. (2025). YOLOv8-AM: YOLOv8 based on effective attention mechanisms for paediatric wrist fracture detection. *IEEE Access, 13*, 44118-44133. https://doi.org/10.1109/ACCESS.2025.3549839

Cui, X., Guan, Z., Morgan, K., Huang, K., & Hammami, A. (2022). Multitiered fresh produce supply chain: The case of tomatoes. *Horticulturae, 8*(12), Article 1204. https://doi.org/10.3390/horticulturae8121204

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R. R., Cheng, M., & Hu, S. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media, 8*(3), 331-368. https://doi.org/10.1007/s41095-022-0271-y

Li, P., Zheng, J., Li, P., Long, H., Li, M., & Gao, L. (2023). Tomato maturity detection and counting model based on MHSA-YOLOv8. *Sensors, 23*(15), Article 6701. https://doi.org/10.3390/s23156701

Li, R., Ji, Z., Hu, S., Huang, X., Yang, J., & Li, W. (2023). Tomato maturity recognition model based on improved YOLOv5 in greenhouse. *Agronomy, 13*(2), Article 603. https://doi.org/10.3390/agronomy13020603

Moreira, G., Magalhães, S. A., Pinho, T., Santos, F. N. D., & Cunha, M. (2022). Benchmark of deep learning and a proposed HSV colour space model for the detection and classification of greenhouse tomato. *Agronomy, 12*(2), Article 356. https://doi.org/10.3390/agronomy12020356

Phoophuangpairoj, R., Ngoenrungrueang, T., & Audomsin, S. (2023). Ripeness classification of a bunch of bananas using a CNN. *2022 International Electrical Engineering Congress (iEECON)*, 180-183. https://doi.org/10.1109/iEECON56657.2023.10126585

Rizzo, M., Marcuzzo, M., Zangari, A., Gasparetto, A., & Albarelli, A. (2023). Fruit ripeness classification: A survey. *Artificial Intelligence in Agriculture, 7*, 44-57. https://doi.org/10.1016/j.aiia.2023.02.004

Su, F., Zhao, Y., Wang, G., Liu, P., Yan, Y., & Zu, L. (2022). Tomato maturity classification based on SE-YOLOv3-MobileNetV1 network under natural greenhouse environment. *Agronomy, 12*(7), Article 1638. https://doi.org/10.3390/agronomy12071638

Tang, C., Chen, D., Wang, X., Ni, X., Liu, Y., Liu, Y., Mao, X., & Wang, S. (2023). A fine recognition method of strawberry ripeness combining Mask R-CNN and region segmentation. *Frontiers in Plant Science, 14*, Article 1211830. https://doi.org/10.3389/fpls.2023.1211830

Wang, C., Han, Q., Li, J., Li, C., & Zou, X. (2024). YOLO-BLBE: A novel model for identifying blueberry fruits with different maturities using the I-MSRCR method. *Agronomy, 14*(4), Article 658. https://doi.org/10.3390/agronomy14040658

Wang, C., Wang, C., Wang, L., Wang, J., Liao, J., Li, Y., & Lan, Y. (2023). A lightweight cherry tomato maturity real-time detection algorithm based on improved YOLOv5n. *Agronomy, 13*(8), Article 2106. https://doi.org/10.3390/agronomy13082106

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv*. https://doi.org/10.1007/978-3-031-72751-1_1

Wang, D., Wang, X., Chen, Y., Wu, Y., & Zhang, X. (2023). Strawberry ripeness classification method in facility environment based on red colour ratio of fruit rind. *Computers and Electronics in Agriculture, 214*, Article 108313. https://doi.org/10.1016/j.compag.2023.108313

Yang, W., Ma, X., & An, H. (2023). Blueberry ripeness detection model based on enhanced detail feature and content-aware reassembly. *Agronomy, 13*(6), Article 1613. https://doi.org/10.3390/agronomy13061613

Yao, H., Liu, Y., Li, X., You, Z., Feng, Y., & Lu, W. (2022). A detection method for pavement cracks combining object detection and attention mechanism. *IEEE Transactions on Intelligent Transportation Systems, 23*(11), 22179-22189. https://doi.org/10.1109/TITS.2022.3177210

Zamri, F. N. M., Gunawan, T. S., Yusoff, S. H., Alzahrani, A. A., Bramantoro, A., & Kartiwi, M. (2024). Enhanced small drone detection using optimised YOLOv8 with attention mechanisms. *IEEE Access, 12*, 90629-90643. https://doi.org/10.1109/ACCESS.2024.3420730

Zeng, T., Li, S., Song, Q., Zhong, F., & Wei, X. (2023). Lightweight tomato real-time detection method based on improved YOLO and mobile deployment. *Computers and Electronics in Agriculture, 205*, Article 107625. https://doi.org/10.1016/j.compag.2023.107625