YOLOv8s-Swin: Enhanced Tomato Ripeness Detection for Smart Agriculture

Jalal Uddin Md Akbar, Syafiq Fauzi Kamarulzaman*
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah
Pekan, 26600, Pahang, Malaysia

Abstract—**Accurate object detection and classification are paramount in precision agriculture for assessing ripeness stages and optimizing yield, particularly for high-value crops like tomatoes. Traditional manual inspection methods are laborious, time-consuming, and error-prone. Furthermore, existing deep learning models often struggle with real-world agricultural challenges such as varying lighting, occlusions from foliage or other fruits, and dense clustering of small objects. To address these limitations and enhance tomato production efficiency and quality in diverse agricultural conditions, this study introduces YOLOv8s-Swin, an advanced object detection model. YOLOv8s-Swin integrates the powerful YOLOv8s architecture with a Swin Transformer module (C3STR) to capture global and local contextual information, crucial for robust small object detection. It also incorporates Focus, Depthwise Convolution (DWconv), Spatial Pyramid Pooling with Contextual Spatial Pyramid Convolution (SPPCSPC), and C2 modules for preserving fine details, reducing computational overhead, enhancing multi-scale feature fusion, and improving high-level semantic feature extraction, respectively. The Wise Intersection over Union (WIoU) loss function is adopted to enhance localization and address convergence issues. Evaluated on a comprehensive tomato image dataset, YOLOv8s-Swin demonstrated superior performance with a mean Average Precision (mAP@0.5) of 88.3%, precision of 84.4%, recall of 79.9%, and an F1-Score of 0.821. This significantly surpasses the base YOLOv8s (84.7% mAP@0.5, 0.795 F1-Score) and other models like Faster R-CNN, SSD, YOLOv4, YOLOv5s, and YOLOv7, all under identical conditions. Maintaining a competitive inference speed of 166.67 FPS, YOLOv8s-Swin offers a robust and efficient solution for AI-driven crop management and sustainable food production.**

Keywords*—*Agricultural automation; attention mechanism; computer vision; smart agriculture; object detection; YOLO; swin transformer

I. Introduction

The agricultural sector is undergoing a profound transformation driven by advancements in computer vision (CV) and deep learning algorithms. Accurate object detection, a key facet of CV, plays a pivotal role in modern precision agriculture, enabling automated tasks such as crop variety identification and ripeness assessment. For high-value crops like tomatoes, optimizing yield and quality within dynamic agricultural environments necessitates precise and timely detection and classification of their ripeness levels [1]. Traditional manual inspection methods are laborious, time-consuming, and error-prone. Convolutional Neural Networks (CNNs) have spearheaded significant progress in this domain, providing robust solutions for extracting hierarchical visual features from images [2]. Among the prominent deep learning architectures, the You Only Look Once (YOLO) series has revolutionized object detection by achieving remarkable speed and accuracy in a single forward pass [3]. Each iteration of the YOLO architecture strives to push the boundaries of both computational efficiency and detection performance. This study focuses on leveraging and enhancing the YOLOv8 architecture, known for its balance of speed and precision, particularly in detecting smaller objects [4].

Despite the efficacy of current deep learning models, challenges persist in real-world agricultural settings. Factors such as varying lighting conditions, occlusions from foliage or other fruits, and dense clustering can significantly impact detection accuracy [5]. Small objects, in particular, often lack sufficient semantic information due to limited pixels, making their accurate detection difficult [6]. To address these challenges, researchers have increasingly explored the integration of attention mechanisms and advanced architectural components into object detection models. Attention mechanisms allow models to selectively focus on the most relevant features in an image, improving feature representation and reducing the influence of irrelevant background noise [7], [8].

Inspired by the success of transformer-based architectures in capturing long-range dependencies and global contextual information in natural language processing and, more recently, in computer vision [9], this study proposes YOLOv8s-Swin, an enhanced object detection model for tomato ripeness detection in smart agriculture. Previous versions of YOLO, including YOLOv4 and YOLOv5s, have shown strong performance in object detection but face significant challenges when it comes to detecting small, occluded, or densely clustered objects in complex environments [25]. These models often struggle with contextual awareness, which is critical for distinguishing between closely packed fruits or handling occlusions from foliage. While YOLOv8s introduced improvements in detection speed and accuracy, particularly for small object detection, it still shares similar limitations to its predecessors. The base YOLOv8s model's focus on local feature extraction and its relatively shallow architectural layers hinder its ability to capture broader contextual information needed for accurate detection in agricultural environments [26]. The model's performance drops when faced with the dense clustering of tomatoes or partial occlusions by other fruits or leaves. To address these issues, YOLOv8s-Swin integrates the Swin Transformer module, which allows for better capture of both local and global context through its self-attention mechanism. This integration, along with additional architectural improvements such as the Focus, Depthwise Convolution (DWconv), SPPCSPC, and C2 modules, enhances the model's ability to handle small and occluded objects while improving feature extraction, computational efficiency, and multi-scale feature fusion. Thus, YOLOv8s-Swin overcomes the limitations of the base YOLOv8s model, providing a more robust and efficient solution for detecting tomatoes at various ripeness stages under diverse agricultural conditions.

The primary contributions of this study are:

- Development of YOLOv8s-Swin, an enhanced architecture for robust tomato ripeness detection. This model integrates the Swin Transformer (C3STR), Focus, DWconv, SPPCSPC, C2 modules, and the WIoU loss function, optimizing feature extraction and localization for challenging agricultural scenes.
- Demonstration of superior performance in classifying tomato ripeness across diverse agricultural conditions. YOLOv8s-Swin achieved an 88.3% mAP@0.5, significantly outperforming the base YOLOv8s (84.7%) and other models, effectively handling occlusions and lighting variations.
- Achieving high accuracy while maintaining real-time inference capabilities. With an FPS of 166.67, YOLOv8s-Swin offers a practical and efficient solution for AI-driven crop management and sustainable food production.

The remainder of this study is organized as follows: Section II provides a comprehensive review of relevant literature on object detection, with a focus on advanced architectures. Section III details the materials and methods used, including data acquisition, preprocessing, and the architectural specifics of the proposed YOLOv8s-Swin model, as well as the experimental setup. Section IV presents and discusses the quantitative and qualitative results, including a comparative analysis with baseline models. Finally, Section V concludes the study by summarizing key findings and outlining directions for future research.

II. Related Work

Object detection has been a cornerstone of computer vision for decades. Early approaches, such as two-stage detectors like R-CNN [10], Fast R-CNN [11], and Faster R-CNN [12], achieved high accuracy by first generating region proposals and then classifying and refining them. While accurate, these methods often struggled with real-time applications due to their computational intensity.

The advent of one-stage detectors, notably the YOLO (You Only Look Once) series, marked a significant shift by performing object localization and classification in a single forward pass, dramatically increasing detection speed [3]. Subsequent YOLO versions, including YOLOv2 [13], YOLOv3 [14], YOLOv5 [15], YOLOv7 [16], and YOLOv8 [4], have continuously pushed the boundaries of speed and accuracy, making them highly suitable for real-time applications. YOLOv8, in particular, introduced advancements in its backbone network and decoupled heads, enhancing its performance, especially for smaller objects.

Small object detection remains a challenging problem due to the limited pixel information available for feature extraction and the lack of sufficient contextual background [6], [27]. To address this, various strategies have been proposed. Attention mechanisms have also gained prominence, allowing models to focus on important features and suppress irrelevant ones, thereby improving representation [7], [8]. Channel attention (e.g., Squeeze-and-Excitation Networks [17]) and spatial attention (e.g., Convolutional Block Attention Module (CBAM) [18]) are widely used variants. Recent works, such as those by Chien et al. [19] and Najihah Muhamad Zamri et al. [20], have demonstrated the benefits of integrating attention mechanisms into YOLO models for improved detection in various applications.

The introduction of Transformer architectures, initially successful in natural language processing, has significantly impacted computer vision. The Swin Transformer [9] is particularly relevant for its hierarchical vision transformer using shifted windows, enabling it to capture global and local contextual information efficiently while maintaining computational feasibility[28]. This makes it a strong candidate for enhancing object detection models, especially for challenging scenarios like small object detection in complex environments. Shi et al. [6] successfully integrated a Swin Transformer module into a YOLOv8 network (STF-YOLO) for small object detection of tea buds, demonstrating significant improvements in accuracy.

In the context of crop ripeness detection, several studies have leveraged YOLO-based models. Li et al. [21] used YOLOv5 for tomato maturity detection in greenhouses, achieving high precision but facing potential limitations in outdoor conditions. Li et al. [22] proposed MHSA-YOLOv8 for tomato grading and counting, which performed well in complex environments but struggled with severe occlusion. Lightweight approaches like those by Zeng et al. [23] and Su et al. [24] emphasized real-time performance but sometimes faltered with dense small targets. While these studies show the promise of YOLO in agriculture, the specific challenges of dense fruit clusters, occlusions, and varied lighting in tomato detection within smart agriculture, combined with the need for robust small object detection, highlight the necessity for further architectural enhancements. Our proposed YOLOv8s-Swin aims to bridge this gap by combining the strengths of YOLOv8s with the contextual awareness of the Swin Transformer and other efficient modules to improve tomato ripeness detection.

III. Materials and Methods

This section details the proposed YOLOv8s-Swin model, its integration of key components, and the comprehensive methodological workflow for enhanced tomato ripeness detection in smart agriculture.

A. Data Acquisition

For this research, the Laboro Tomato dataset (https://github.com/laboroai/LaboroTomato) was utilized. This dataset contains images of tomatoes at three ripeness stages (fully-ripened, half-ripened, and green), reflecting diverse agricultural conditions and camera variations. As shown in Fig. 1, fully ripened tomatoes exhibit >90% red, half-ripened 30-89% red, and green 0-30% red coloration, crucial for accurate ripeness classification.

B. Data Pre-processing and Preparation

Raw images underwent rigorous preprocessing: auto-orientation, uniform resizing to 640×640 pixels, class modification (see Table I), and null annotation filtering. The dataset was split into training (92%, 2001 images), validation (4%, 86 images), and test (4%, 87 images) sets. Extensive data augmentation via Roboflow (see Fig. 2) was applied, including random flips, rotations, shears, blur, and noise, to simulate diverse agricultural conditions and enhance generalization.

TABLE I. 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

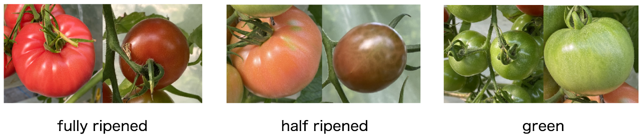

Fig. 1. Dataset class.

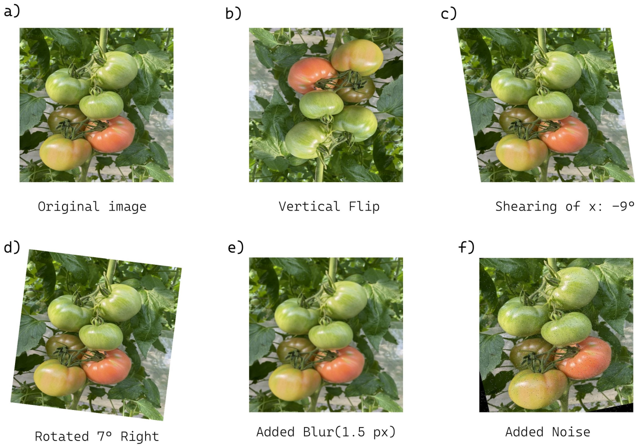

Fig. 2. Data augmentation techniques applied to tomato images; (a) Original image, (b) Vertical Flip, (c) Shearing of X: -9°, (d) Rotated 7° Right, (e) Added Blur (1.5px), (f) Added Noise.

C. Proposed Method: YOLOv8s-Swin Architecture

YOLOv8s-Swin is an advanced object detection architecture building on YOLOv8s, integrating a Swin Transformer module and other components. This enhances focus and contextual information capture, improving accuracy for small and occluded objects. The architecture follows a typical Backbone, Neck, and Head structure, with modifications to feature processing (see Fig. 3).

1) Swin Transformer Module (C3STR): The core innovation is the C3STR module (see Fig. 4), a Swin Transformer integration that mitigates feature loss in deep networks by establishing global dependencies via self-attention [9]. Integrated into the YOLOv8s backbone and neck, C3STR enhances semantic information and representation for small objects. It comprises Window/Shifted Window Multi-Headed Self-Attention (W-MSA/SW-MSA) and Multi-Layer Perceptron (MLP) with internal residual connections. Its self-attention mechanism is:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{1}$$

C3STR controls computations within local windows, enabling cross-window information interaction while reducing complexity, allowing adaptive feature interaction and capturing crucial contextual information.

2) Additional Modules: Focus, DWconv, SPPCSPC, and C2: To optimize YOLOv8s-Swin, efficient modules are incorporated:

- Focus Module: Placed at the input, it increases channel dimensions while reducing spatial ones, preserving fine details for small objects.
- Depthwise Convolution (DWconv): Performs independent channel convolutions, reducing parameters and computation while maintaining high performance.
- Spatial Pyramid Pooling with Contextual Spatial Pyramid Convolution (SPPCSPC): Captures multi-scale feature information efficiently via pooling and convolution.
- C2 Module: A refined C3 version for enhanced high-level semantic feature extraction with improved memory and inference speed.

These modules, integrated into the YOLOv8s backbone and neck, enable accurate tomato ripeness detection in complex agricultural conditions.

3) Loss Function (WIoU): To address YOLOv8's convergence issues and enhance localization, the Wise Intersection over Union (WIoU) loss function is adopted. Unlike traditional IoU, WIoU considers positional relationships and geometric factors, mitigating the negative impact of low-quality examples and reducing centroid distance emphasis for well-overlapping frames. The WIoUv1 model's calculation is:

$$L_{WIOU \cdot wl} = R_{WIoU} L_{IOU} \tag{2}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gl})^2 + (y - y_{gl})^2}{(W_g^2 + H_g^2)^*}\right) \tag{3}$$

$$L_{IOU} = 1 - IOU = 1 - \frac{W_i H_i}{S_u} \tag{4}$$

Here, $R_{\mathrm{WIoU}} \in [1, e)$ amplifies L_{IoU} for normal anchor frames but reduces it for high-quality ones, focusing learning

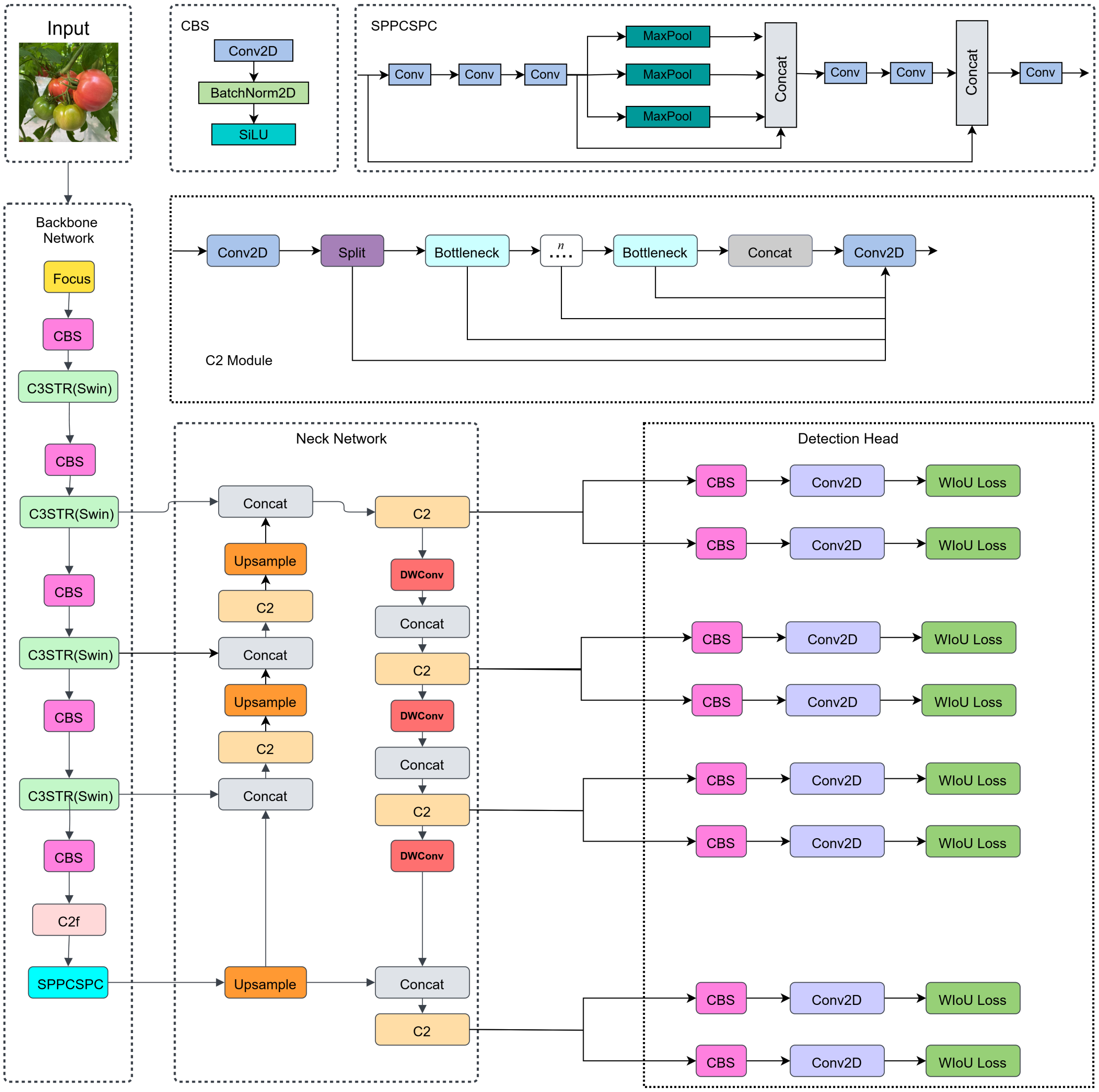

Fig. 3. The architecture of YOLOv8s-Swin for tomato ripeness detection.

on better examples for improved generalization and stability. S_u is the union area.

D. Experimental Setup

Experiments were conducted on an NVIDIA GeForce RTX 3060 GPU (16 GB VRAM) with an AMD Ryzen 5 processor, running Ubuntu 24.04 LTS and PyTorch 2.3.0. Training used 640×640 pixel images with a batch size of 8. Key hyper-parameters included: 150 epochs, initial learning rate of 0.01, momentum 0.937, weight decay 0.0005, and data augmentation parameters (mixup 0.15, copy paste 0.3).

IV. Results and Discussion

A. Evaluation Metrics

The performance of the YOLOv8s-Swin model for tomato ripeness detection was quantitatively assessed using several standard object detection metrics:

1) Precision (P): The proportion of correctly identified positive samples among all detected positives. The precision metric is calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

In this context, TP refers to true positives, representing instances that were correctly identified as positive. Conversely, FP refers to false positives, which are instances incorrectly classified as positive. Attaining a higher precision value signifies a reduction in false alarms within the detection process, leading to a more reliable model.

2) Recall (R): The ratio of correctly identified positive samples to the total number of actual positive samples. The recall metric is defined mathematically as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

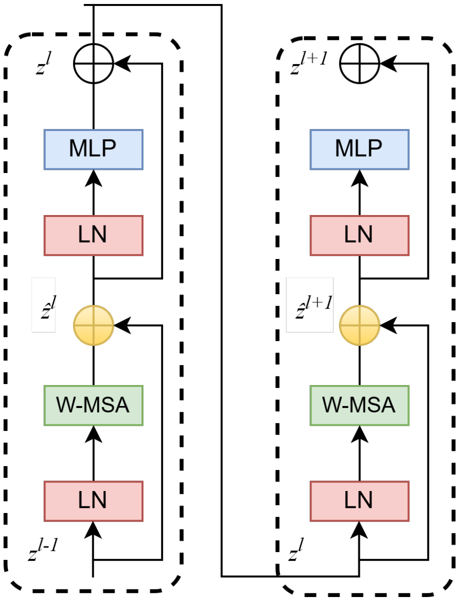

Fig. 4. C3STR architecture diagram.

Here, TP corresponds to true positives, referring to correctly identified positive cases. FN, on the other hand, stands for false negatives, which are genuine positive instances that the model overlooked. Achieving a higher recall value indicates a reduction in missed relevant instances, thereby ensuring a more complete detection capability.

3) mAP (mean average precision): Calculated at an Intersection over Union (IoU) threshold of 0.5 (mAP@0.5) and also across multiple IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95). It represents the average of the Average Precision (AP) for all classes, derived from the area under the precision-recall curve.

4) F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the model's accuracy.

5) FPS (Frames Per Second): Measures the number of images the model can process per second at a batch size of 1, indicating the model's detection speed.

B. Performance of YOLOv8s-Swin

The proposed YOLOv8s-Swin model was rigorously evaluated on the Laboro Tomato dataset, encompassing fully ripe, partially ripe, and unripe tomatoes in diverse agricultural environments. The model demonstrated strong performance across all evaluation metrics, showcasing its effectiveness in accurately detecting and classifying tomato ripeness stages.

Our YOLOv8s-Swin model achieved an impressive mAP@0.5 of 88.3%, a precision of 84.4%, and a recall of 79.9%. The F1-Score for the model was 0.821. For individual classes, the precision-recall curve (see Fig. 5) highlights mAP@0.5 values of 0.893 for 'fully_ripened', 0.882 for 'green', and 0.873 for 'half_ripened' tomatoes, indicating consistent performance across ripeness stages. The F1-Confidence Curve (see Fig. 6) shows an optimal F1-Score of 0.84 for all classes at a confidence threshold of 0.456.

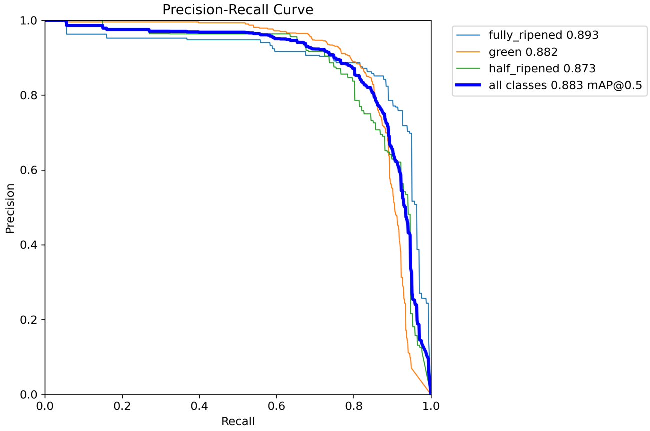

Fig. 5. Precision-recall curve of proposed YOLOv8s-Swin.

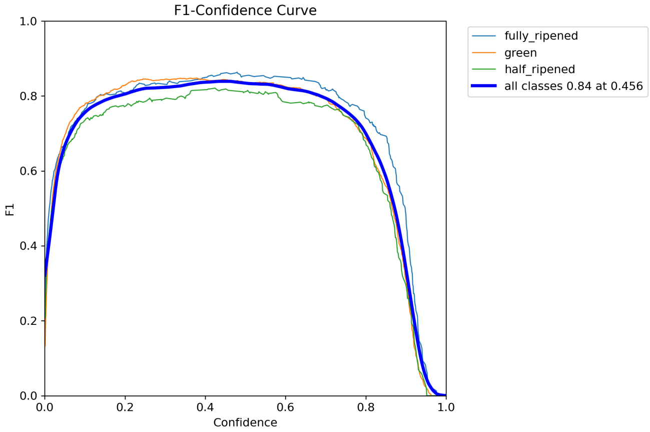

Fig. 6. F1-Confidence curve of proposed YOLOv8s-Swin.

The confusion matrix (see Fig. 7) further confirms the model's strong classification ability. It shows high true positive counts for 'fully_ripened' (144), 'green' (476), and 'half_ripened' (134) predictions, indicating a low rate of misclassification. Furthermore, the training and validation loss curves for bounding box regression, classification, and the overall loss (see Fig. 8) exhibit a consistent and steady decline over training epochs, demonstrating effective learning and robust generalization to unseen data.

C. Comparative Performance Analysis

To thoroughly assess the efficacy and advantages of our proposed YOLOv8s-Swin algorithm, a comprehensive comparative analysis was conducted against several prominent object detection models. These models include traditional two-stage detectors like Faster R-CNN, one-stage detectors such as SSD, and various iterations of the YOLO series, namely YOLOv4, YOLOv5s, YOLOv7, and the base YOLOv8s model. Crucially, all comparative experiments were performed under strictly controlled and identical conditions, utilizing the same hardware setup (NVIDIA GeForce RTX 3060 GPU), the Laboro Tomato dataset, consistent data augmentation methodologies, and a consistent training regimen of 150 epochs, with the optimal

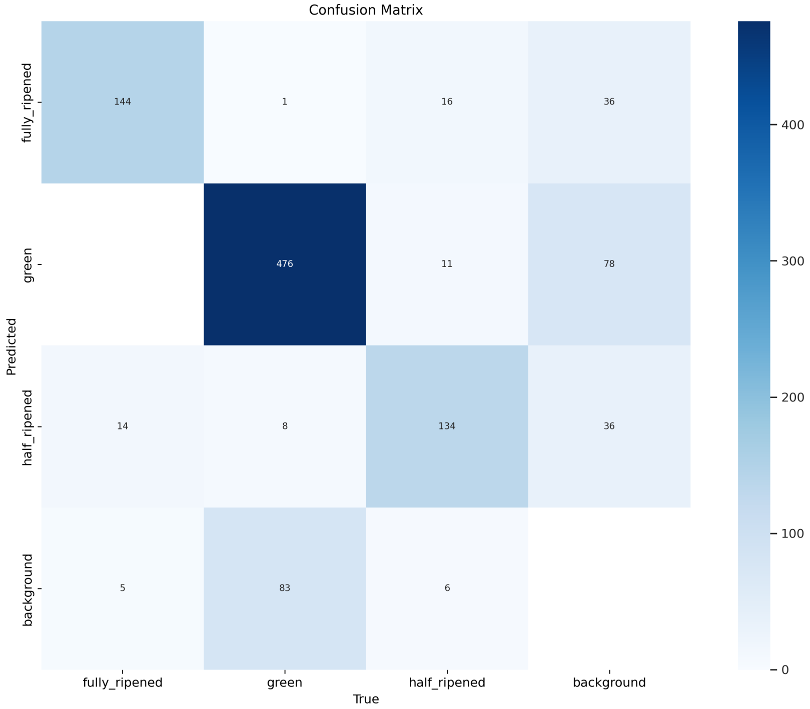

Fig. 7. Confusion matrix of proposed YOLOv8s-Swin.

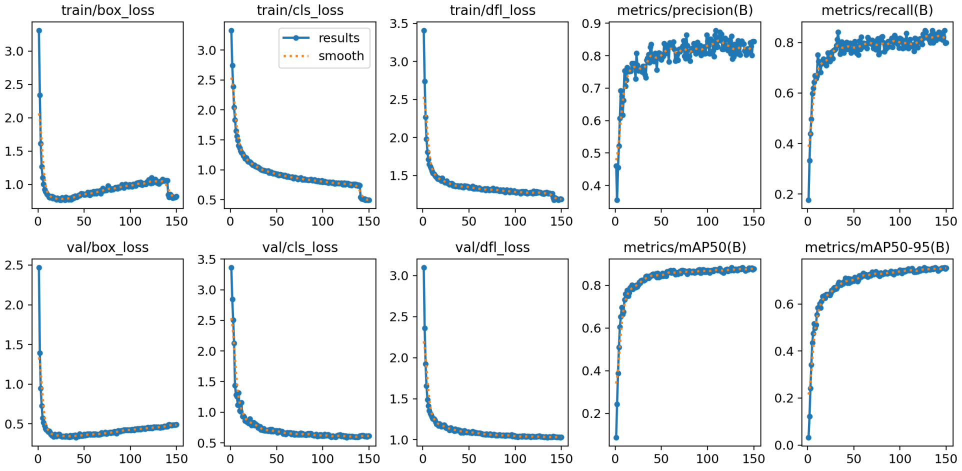

Fig. 8. Visualization of different loss curves of proposed YOLOv8s-Swin model over epochs.

results selected for evaluation.

As presented in Table II, the proposed YOLOv8s-Swin model demonstrates significant improvements across key performance metrics, particularly when directly compared to its base counterpart, YOLOv8s. YOLOv8s-Swin achieved a notably higher mAP@0.5 of 88.3% compared to the base YOLOv8s's 84.7%, representing a substantial improvement in overall detection accuracy under identical experimental settings. Furthermore, YOLOv8s-Swin's precision is 84.4%, an increase from YOLOv8s's 78.3%, indicating its superior accuracy in identifying true positive instances. The recall for YOLOv8s-Swin also saw an improvement, reaching 79.9% compared to YOLOv8s's 78.2%, meaning it is more effective at capturing a higher proportion of actual positive cases. The F1-Score, a crucial balanced measure of precision and recall, also significantly improved from 0.795 for YOLOv8s to 0.821 for YOLOv8s-Swin. These advancements in detection accuracy, precision, recall, and F1-Score are a direct result of the integrated architectural enhancements.

The superior performance of YOLOv8s-Swin is primarily due to the strategic integration of several advanced modules into the YOLOv8 framework. Specifically, the Swin Transformer module significantly enhances the model's ability to capture global and local contextual information, which is

TABLE II. Comprehensive Model Comparison for Tomato Ripeness Detection

Network	P (%)	R (%)	mAP@0.5 (%)	F1-Score	FPS
Faster R-CNN	82.0	78.0	86.0	0.780	10
SSD	78.6	75.0	78.0	0.740	40
YOLOv4	74.0	70.0	76.0	0.719	55
YOLOv5s	78.0	73.5	82.0	0.762	110
YOLOv7	78.3	75.0	83.0	0.779	120
YOLOv8s	78.3	78.2	84.7	0.795	160
YOLOv8s-Swin (Proposed)	**84.4**	**79.9**	**88.3**	**0.821**	**166.67**

crucial for accurately detecting small and partially occluded tomatoes in complex agricultural scenes. Additionally, the inclusion of the Focus module aids in preserving fine-grained details vital for small objects, Depthwise Convolution (DWconv) reduces computational overhead while maintaining feature richness, SPPCSPC enhances multi-scale feature fusion, and the C2 module improves high-level semantic feature extraction. These combined modules, along with the robust WIoU loss function, ensure that critical information is effectively utilized during the learning process, leading to more reliable and robust feature representations, particularly beneficial for accurately recognizing subtle differences between various stages of tomato maturity.

Beyond the direct comparison with YOLOv8s, YOLOv8s-Swin exhibits remarkable performance against all other models tested under identical experimental settings. It achieves the highest mAP@0.5 of 88.3%, significantly surpassing traditional two-stage detectors like Faster R-CNN (86.0%) and SSD (78.0%), as well as earlier YOLO iterations such as YOLOv4 (76.0%), YOLOv5s (82.0%), and YOLOv7 (85.0%). The F1-Score of 0.821 is also the highest among all compared models, underscoring its superior balance between precision and recall for robust real-world performance.

In terms of detection speed, YOLOv8s-Swin maintains a highly competitive FPS rate of 166.67, which is marginally higher than the base YOLOv8s model's 160 FPS. This satisfies the stringent requirements for real-time detection in agricultural applications. In contrast, older models like Faster R-CNN (10 FPS) and SSD (40 FPS) fall significantly short of real-time capabilities. This critical balance between high accuracy and efficient processing makes YOLOv8s-Swin a highly universal and practical solution for smart agriculture, particularly suitable for deployment on edge devices with limited processing capabilities without substantial hardware investments.

D. Qualitative Analysis

To visually demonstrate the effectiveness of the proposed YOLOv8s-Swin model, Fig. 9 presents several detection examples across various agricultural conditions, including instances with varying lighting, occlusions, and different ripeness stages. The model consistently exhibits strong performance in identifying and classifying tomatoes at their respective ripeness levels. For instance, Fig. 9(a), Fig. 9(b), and Fig. 9(f) clearly illustrate the model's capability to accurately detect and categorize green (unripe) tomatoes, even when they are densely clustered or partially obscured by surrounding foliage. This highlights the model's resilience to common real-world challenges in agricultural settings.

Furthermore, Fig. 9(c) and Fig. 9(e) showcase the precise detection of fully-ripened tomatoes, with the bounding boxes and associated confidence scores, indicating a high degree of accuracy in classification. This performance is crucial for timely harvesting and yield optimization. Fig. 9(d) specifically demonstrates the model's nuanced ability to differentiate between half-ripened and fully-ripened tomatoes, underscoring its capacity to discern subtle visual cues that distinguish various stages of maturity. Collectively, these qualitative results visually confirm the robustness and practical applicability of the YOLOv8s-Swin model in diverse and challenging agricultural environments, reinforcing its strong quantitative performance metrics through accurate bounding box localization and precise ripeness classification under real-world conditions.

E. Discussion

Accurate detection and classification of crop ripeness stages are indispensable for informed decision-making and optimizing yield management in precision agriculture. Our proposed YOLOv8s-Swin model significantly advances the state-of-the-art for tomato ripeness detection, demonstrating a new benchmark for object detection and classification in agricultural applications, specifically for tomatoes.

The superior performance of YOLOv8s-Swin, particularly its notably higher mAP@0.5, precision, recall, and F1-Score compared to baseline models like YOLOv8s and other prominent architectures, is primarily attributable to its innovative architectural enhancements. The strategic integration of the Swin Transformer module (C3STR) proved crucial, allowing the model to better capture global and local contextual information. This enhanced contextual awareness is particularly beneficial for accurately detecting small and partially occluded tomatoes in complex agricultural scenes. Furthermore, the inclusion of the Focus module aids in preserving fine-grained details, while Depthwise Convolution (DWconv) reduces computational overhead without sacrificing feature richness. The SPPCSPC module enhances multi-scale feature fusion, and the C2 module improves high-level semantic feature extraction by being a refined version of C3. These combined architectural modifications, coupled with the robust Wise Intersection over Union (WIoU) loss function, which mitigates the negative impact of low-quality examples and enhances localization, ensure that critical information is effectively utilized during the learning process. This leads to more reliable and robust feature representations, particularly beneficial for accurately recog-

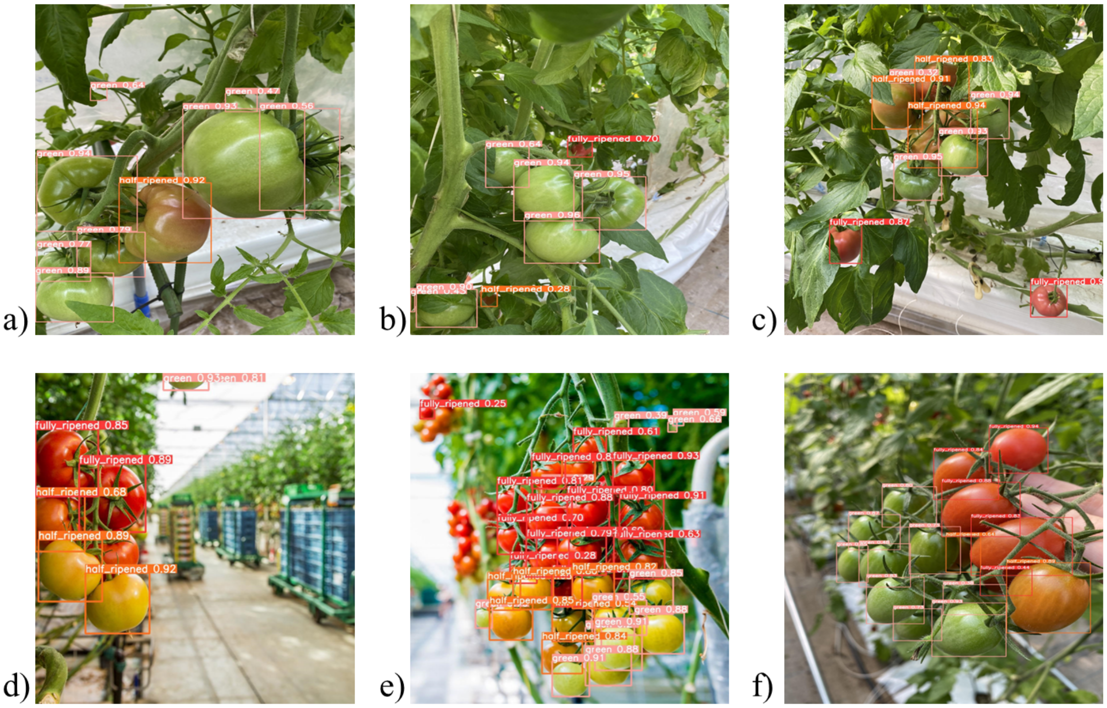

Fig. 9. YOLOv8s-Swin's robust detection capabilities across diverse smart agriculture conditions, including: (a) fruits occluded by other fruits, (b) fruits occluded by foliage, (c) combined occlusion by foliage and fruits, (d) detection in high-density fruit environments, (e) accurate detection in densely packed fruits, and (f) high-density fruit clustering detection.

nizing subtle differences between various stages of tomato maturity.

Beyond its significant accuracy improvements, YOLOv8s-Swin maintains a highly competitive inference speed of 166.67 FPS, which is practical for real-time applications in agriculture. This speed, combined with the model's moderate computational requirements, facilitates efficient deployment on edge devices commonly found in agricultural settings, thereby minimizing the need for substantial hardware investments. The model's adaptability also suggests potential for customization across various platforms, enhancing scalability for diverse agricultural operations.

Despite the high accuracy achieved, certain challenges persist. The model may still struggle with severely occluded tomatoes or distinguishing them from background foliage in extremely dense clusters. Future research will focus on several key areas to build upon these findings. Integrating multi-modal data sources, such as infrared or depth images, could significantly enhance the model's robustness against challenging occlusions and lighting variations. Furthermore, exploring advanced optimization techniques, including neural architecture search, could lead to even more tailored and efficient designs for this specific task. Deployment-specific optimizations, such as quantization-aware training, will be pursued to further reduce computational requirements, enabling broader and more efficient use in resource-constrained environments.

V. Conclusion

This study successfully introduced YOLOv8s-Swin, an enhanced object detection model for robust tomato ripeness detection within smart agriculture. A primary contribution lies in the innovative integration of the Swin Transformer module, Focus, Depthwise Convolution (DWconv), Spatial Pyramid Pooling with Contextual Spatial Pyramid Convolution (SPPCSPC), and C2 modules into the YOLOv8s framework. These architectural enhancements proved highly effective, significantly improving the model's ability to detect small, occluded objects and handle complex agricultural environments. The model achieved a mean Average Precision (mAP@0.5) of 88.3%, precision of 84.4%, recall of 79.9%, and an F1-Score of 0.821, surpassing the base YOLOv8s (84.7% mAP@0.5, 78.3% precision, 78.2% recall, 0.795 F1-Score) and other models like Faster R-CNN, SSD, YOLOv4, YOLOv5s, and YOLOv7, all trained under identical conditions. These improvements are attributed to the model's ability to capture both local and global context, addressing the limitations of previous YOLO versions, which struggled with occlusions, lighting variations, and dense clustering of objects. Additionally, YOLOv8s-Swin maintains a competitive inference speed of 166.67 FPS, demonstrating its applicability in real-time applications for AI-driven crop management and sustainable food production. While the model has demonstrated impressive performance, future work will focus on optimizing it for deployment in more resource-constrained environments, integrating multimodal data sources like infrared and depth sensing to further enhance accuracy under challenging conditions, and exploring the applicability of the model to other crops. The advancements presented here significantly contribute to the field of smart agriculture by providing an efficient and scalable solution for detecting ripeness stages in tomatoes, paving the way for future research and development in AI-driven agricultural automation.

Acknowledgment

We sincerely thank the UMPSA Postgraduate Research Grants Scheme (PGRS) for their support through the Open Access funding, which made this research possible.

References

[1] M. Rizzo, M. Marcuzzo, A. Zangari, A. Gasparetto, and A. Albarelli, “Fruit ripeness classification: A survey,” *Artificial Intelligence in Agriculture*, vol. 7, pp. 44–57, 2023. doi: 10.1016/j.aiia.2023.02.004.

[2] J. U. Md. Akbar, S. F. Kamarulzaman, A. J. Md. Muzahid, M. A. Rahman, and M. Uddin, “A Comprehensive Review on Deep Learning Assisted Computer Vision Techniques for Smart Greenhouse Agriculture,” *IEEE Access*, vol. 12, pp. 4485–4522, 2024. doi: 10.1109/ACCESS.2024.3349418.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016.

[4] S. Ma, H. Lu, J. Liu, Y. Zhu and P. Sang, “LAYN: Lightweight Multi-Scale Attention YOLOv8 Network for Small Object Detection,” in IEEE Access, vol. 12, pp. 29294-29307, 2024, doi: 10.1109/ACCESS.2024.3368848.

[5] R. Li, Z. Ji, S. Hu, X. Huang, J. Yang, and W. Li, “Tomato Maturity Recognition Model Based on Improved YOLOv5 in Greenhouse,” *Agronomy*, vol. 13, no. 2, Art. no. 603, 2023. doi: 10.3390/agronomy13020603.

[6] M. Shi, D. Zheng, T. Wu, W. Zhang, R. Fu, and K. Huang, “Small object detection algorithm incorporating swin transformer for tea buds,” *PLoS ONE*, vol. 19, no. 3, pp. 1–25, 2024. doi: 10.1371/journal.pone.0299902.

[7] M.-H. Guo *et al.*, “Attention mechanisms in computer vision: A survey,” *Computational Visual Media*, vol. 8, no. 3, pp. 331-368, 2022. doi: 10.1007/s41095-022-0271-y.

[8] H. Yao, Y. Liu, X. Li, Z. You, Y. Feng, and W. Lu, “A Detection Method for Pavement Cracks Combining Object Detection and Attention Mechanism,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22179–22189, 2022. doi: 10.1109/TITS.2022.3177210.

[9] Z. Liu *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012-10022.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580-587.

[11] J. Li, X. Liang, S. Shen, T. Xu, J. Feng and S. Yan, “Scale-Aware Fast R-CNN for Pedestrian Detection,” in *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985-996, April 2018, doi: 10.1109/TMM.2017.2759508.

[12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[13] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017.

[14] Q. -C. Mao, H. -M. Sun, Y. -B. Liu and R. -S. Jia, “Mini-YOLOv3: Real-Time Object Detector for Embedded Applications,” in IEEE Access, vol. 7, pp. 133529-133538, 2019, doi: 10.1109/ACCESS.2019.2941547.

[15] C. Wang *et al.*, “A Lightweight Cherry Tomato Maturity Real-Time Detection Algorithm Based on Improved YOLOV5n,” *Agronomy*, vol. 13, no. 8, Art. no. 2106, 2023. doi: 10.3390/agronomy13082106.

[16] Y. Wu *et al.*, “An improved YOLOv7 network using RGB-D multi-modal feature fusion for tea shoots detection,” *Computers and Electronics in Agriculture*, vol. 216, pp. 108541, 2024. doi: 10.1016/j.compag.2023.108541.

[17] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in *Proc. Eur. Conf. Computer Vision (ECCV)*, Munich, Germany, Sep. 2018.

[19] C. -T. Chien, R. -Y. Ju, K. -Y. Chou, E. Xieerke and J. -S. Chiang, “YOLOv8-AM: YOLOv8 Based on Effective Attention Mechanisms for Pediatric Wrist Fracture Detection,” in *IEEE Access*, vol. 13, pp. 52461-52477, 2025, doi: 10.1109/ACCESS.2025.3549839.

[20] F. N. M. Zamri *et al.*, “Enhanced Small Drone Detection Using Optimized YOLOv8 With Attention Mechanisms,” *IEEE Access*, vol. 12, pp. 90629–90643, 2024. doi: 10.1109/ACCESS.2024.3420730.

[21] R. Li *et al.*, “Tomato Maturity Recognition Model Based on Improved YOLOv5 in Greenhouse,” *Agronomy*, vol. 13, no. 2, Art. no. 603, 2023. doi: 10.3390/agronomy13020603.

[22] P. Li *et al.*, “Tomato Maturity Detection and Counting Model Based on MHSA-YOLOv8,” *Sensors*, vol. 23, no. 15, Art. no. 6701, 2023. doi: 10.3390/s23156701.

[23] T. Zeng, S. Li, Q. Song, F. Zhong, and X. Wei, “Lightweight tomato real-time detection method based on improved YOLO and mobile deployment,” *Computers and Electronics in Agriculture*, vol. 205, pp. 107625, 2023. doi: 10.1016/j.compag.2023.107625.

[24] F. Su *et al.*, “Tomato Maturity Classification Based on SE-YOLOv3-MobileNetV1 Network under Nature Greenhouse Environment,” *Agronomy*, vol. 12, no. 7, Art. no. 1638, 2022. doi: 10.3390/agronomy12071638.

[25] Y. Gao, Z. Li, Y. Wang, and S. Zhu, “A Novel YOLOv5_ES based on lightweight small object detection head for PCB surface defect detection,” *Scientific Reports*, vol. 14, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-74368-7.

[26] L. Shen, B. Lang and Z. Song, “DS-YOLOv8-Based Object Detection Method for Remote Sensing Images,” in *IEEE Access*, vol. 11, pp. 125122-125137, 2023, doi: 10.1109/ACCESS.2023.3330844.

[27] J. U. M. Akbar, S. Fauzi Kamarulzaman and E. H. Tusher, “Plant Stem Disease Detection Using Machine Learning Approaches*,” *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-8, doi: 10.1109/ICCCNT56998.2023.10307074.

[28] J. Jumadi and J. U. M. Akbar, “Hybrid GRU-KAN model for energy consumption prediction in commercial building cooling,” International Journal of Advanced Computing and Informatics(IJACI), vol. 1, no. 2, pp. 69–78, Jun. 2025, doi: 10.71129/ijaci.v1.i2.pp69-78.