EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

YOLO-Lychee: A YOLOv8s-Based Detector for Lychee Growth-Stage Detection under Natural Orchard Conditions

Thi-Nguyen Nguyen[1], Manh-Tuan Do[1], Dinh-Thai Kim[2,*], Tuan-Minh Nguyen[2], Huyen-Trang Nguyen[1]

[1]Faculty of Electrical and Electronics Engineering, Viet-Hung Industrial University, 100000, Vietnam
[2]International School, Vietnam National University Hanoi, Hanoi, 100000, Vietnam

Abstract

Accurate lychee growth-stage detection in natural orchards is challenging because blossoms and fruits are often small, densely clustered, partially occluded, and visually similar to surrounding foliage. This study proposes YOLO-Lychee, a YOLOv8s-based detector for four developmental stages: blossom, young fruit, green fruit, and ripe fruit. The dataset contains 1,145 original orchard images and 19,422 annotated instances collected in Hai Duong and Bac Giang provinces, Vietnam. To reduce blossom-class imbalance, exposure-based augmentation was applied only to blossom samples in the training subset, resulting in 1,324 images and 20,261 annotations, while the validation and test sets remained unchanged. YOLO-Lychee replaces the original SPPF module with a Spatial and Channel Cross-Transformer module, incorporates a Context Augmentation Module in the neck, and uses EIoU instead of the default CIoU loss for bounding-box regression. Across six random seeds, YOLO-Lychee achieved a Precision of $83.02 \pm 2.38\%$, Recall of $82.19 \pm 1.87\%$, mAP_{50} of $88.42 \pm 0.38\%$, and $mAP_{50:95}$ of $72.76 \pm 0.45\%$. Compared with recent YOLO-family detectors under the same protocol, the proposed model obtained the highest mAP_{50} and competitive $mAP_{50:95}$ while maintaining real-time inference. Ablation results confirm the complementary contributions of SC3T, CAM, and EIoU, whereas qualitative analysis shows that blossom detection remains the most challenging case. These results demonstrate that YOLO-Lychee is a practical empirical baseline for lychee growth-stage detection and vision-assisted orchard monitoring.

Received on 09 March 2026; accepted on 20 July 2026; published on 27 July 2026

Keywords: Precision agriculture, Deep learning, Object detection, Lychee growth-stage detection, YOLOv8

doi:10.4108/airo.12140

1. Introduction

Lychee is an economically important subtropical fruit crop in Vietnam, particularly in major production areas such as Bac Giang and Hai Duong provinces. Vietnam is among the major lychee-producing countries in Asia, with a cultivation area of 58,340 hectares and an annual production of 380,600 tons reported in 2018 [1]. The development of the Vietnamese lychee industry is closely associated with cultivar selection, flower induction, harvest-season extension, cultivation management, compliance with VietGAP and GlobalGAP standards, and postharvest technologies for domestic and export markets [1]. In this context, accurate monitoring of lychee growth stages is important for orchard management, harvest planning, and quality control.

Robotic harvesting and vision-based orchard monitoring systems have attracted increasing attention because they can reduce labor dependence, improve harvesting consistency, and support data-driven crop management [2]. Fruit detection is a fundamental perception task in such systems. A reliable detector must identify fruits or fruit clusters under natural orchard conditions, where illumination variation, leaf and branch occlusion, background similarity, dense clusters, and scale variation frequently occur [3]. These

*Corresponding author. Email: thaikd@vnu.edu.vn

challenges are particularly evident in lychee orchards because blossoms and fruits are often small, clustered, partially occluded, and visually similar to surrounding foliage.

Deep learning has become a widely used approach for agricultural object detection, fruit recognition, disease identification, and yield estimation. Previous studies have reviewed the role of sensors and computer vision systems in fruit detection and emphasized the effects of illumination, occlusion, and object scale on detection performance [4]. Recent studies have further demonstrated that convolutional neural networks, attention mechanisms, feature-pyramid structures, and YOLO-based detectors can improve both accuracy and inference efficiency in precision-agriculture applications. In addition, related works on artificial intelligence, robotic perception, pest detection, and smart-agriculture systems have highlighted the growing importance of visual perception models for practical agricultural automation and decision support [5–7].

Despite these advances, lychee growth-stage detection remains less studied than the detection of more common fruits such as apples, oranges, grapes, tomatoes, and pomegranates. Many existing fruit-detection studies focus on mature fruit localization or yield estimation, whereas fewer works address multi-stage detection from blossom to ripe fruit. This task is challenging because lychee developmental stages exhibit large differences in size, color, texture, and spatial distribution. Blossom regions are small and sparse, young fruits are densely clustered, green fruits have strong visual similarity to leaves, and ripe fruits may be affected by strong outdoor illumination and partial occlusion. Therefore, lychee growth-stage detection requires a model capable of capturing contextual and multi-scale features for small and visually ambiguous targets.

This study develops a YOLOv8s-based detector, termed YOLO-Lychee, for identifying lychees at four developmental stages: blossom, young fruit, green fruit, and ripe fruit. A stage-specific dataset was collected from commercial lychee orchards in Hai Duong and Bac Giang provinces, Vietnam. The originally acquired dataset comprised 1,145 images and 19,422 annotated instances. To improve the representation of the under-represented blossom class during training, exposure-based augmentation was applied exclusively to blossom images in the training subset. This procedure increased the experimental corpus to 1,324 images and 20,261 annotations. The validation and test subsets were kept unchanged to prevent augmentation leakage and to preserve their original data distributions for model selection and final evaluation.

The proposed detector integrates the Spatial and Channel Cross-Transformer (SC3T) module into the backbone to enhance spatial–channel feature representation, and introduces the Context Augmentation Module (CAM) into the neck to aggregate contextual information from multiple receptive fields. In addition, the Efficient Intersection over Union (EIoU) loss is adopted for bounding-box regression to improve localization accuracy. The model is evaluated against recent YOLO-family detectors and further analyzed through module ablation, CAM fusion comparison, loss-function ablation, class-wise evaluation, efficiency analysis, and qualitative visualization.

The main contributions of this study are summarized as follows:

(1) A stage-specific lychee dataset is constructed under natural orchard conditions in Vietnam, covering blossom, young fruit, green fruit, and ripe fruit.

(2) A YOLOv8s-based detector, named YOLO-Lychee, is developed by replacing the original SPPF module with SC3T, introducing CAM into the neck, and adopting EIoU to improve detection of small, dense, and visually ambiguous lychee targets.

(3) Extensive quantitative, ablation, efficiency, class-wise, and qualitative analyses are conducted to evaluate the proposed detector against recent YOLO-family models under realistic orchard conditions.

The proposed framework is intended to provide an empirical baseline for lychee growth-stage detection and a practical reference for vision-based orchard monitoring. Although further validation across larger datasets, additional orchards, and diverse environmental conditions is still required, the results demonstrate the feasibility of applying YOLO-based object detection to lychee growth-stage analysis in precision agriculture.

The remainder of this paper is organized as follows. Section 2 reviews related studies on deep learning-based fruit detection and lightweight agricultural object detection. Section 3 describes the lychee dataset, YOLOv8s baseline, proposed model components, and evaluation metrics. Section 4 reports the experimental results, comparative analysis, ablation studies, and qualitative evaluation. Section 5 concludes the paper and discusses limitations and future work.

2. Related Work

Deep learning-based object detection has been widely applied to agricultural perception tasks, including fruit localization, maturity assessment, disease identification, yield estimation, and robotic harvesting. Two-stage detectors, such as Faster Region-based Convolutional Neural Network (Faster R-CNN) [8], are known for their strong localization capability in complex visual scenes. In fruit detection, Faster R-CNN with Residual Network-50 (ResNet-50) has been used for coconut maturity-stage detection and achieved a mean Average Precision (mAP) of 0.894, outperforming Single Shot MultiBox Detector (SSD) and YOLOv3 [9]. Bargoti and Underwood applied Faster R-CNN to orchard fruit detection and obtained F1-scores above 0.9 for apples and mangoes, demonstrating the effectiveness of deep convolutional detectors under natural orchard conditions [10]. Other comparative studies have also shown that Faster R-CNN with InceptionV2 can achieve competitive fruit detection and counting performance under illumination variation and partial occlusion [11]. Although two-stage detectors provide accurate localization, their relatively high computational cost makes them less suitable for real-time orchard monitoring and robotic harvesting systems.

One-stage detectors, particularly YOLO-family models, have become increasingly popular in precision agriculture because of their favorable balance between detection accuracy and inference speed. YOLOv4 has been used for orange detection and yield estimation under different illumination conditions, achieving a precision of 91.23% and an mAP of 90.8% [12]. Several YOLO versions have also been evaluated for wild blueberry maturity-stage detection, where YOLOv4 achieved an mAP of 79.79% for three maturity classes and 88.12% for two classes [13]. Lightweight and pruned YOLO variants have further been investigated for real-time agricultural deployment. A pruned YOLOv5s model was proposed for grape-cluster tracking, maintaining an accuracy of 84.9% and an mAP of 82.3% [14]. Similarly, a channel-pruned YOLOv5s model achieved a precision of 95.8% and an F1-score of 91.5% for apple fruitlet detection before thinning [15]. These studies indicate that YOLO-based detectors are practical for agricultural applications where both accuracy and computational efficiency are required.

Recent developments in the YOLO family have further improved real-time detection through more efficient architectures, training strategies, and feature-representation designs. YOLOv8 has been widely adopted as a practical baseline for lightweight object detection because of its implementation efficiency and strong speed–accuracy trade-off [16]. YOLOv9 introduces programmable gradient information to improve information preservation during training [17], while YOLOv10 focuses on real-time end-to-end detection [18]. YOLO11 has been empirically evaluated against earlier YOLO generations [19], and YOLOv12 emphasizes attention-centric real-time detection [20]. More recently, YOLO26 has been introduced as a unified real-time vision model [21]. These models provide relevant baselines for evaluating agricultural object detectors, especially when inference time, model size, FLOPs, and deployment feasibility are considered.

In fruit detection, recent studies have combined YOLO detectors with lightweight modules, pruning strategies, and feature-fusion mechanisms to improve performance in complex orchard environments. An improved YOLOv8 model using ShuffleNetv2 and Ghost modules was developed for real-time multi-stage apple-fruit detection and achieved an mAP of 91.4% [22]. YOLOv5s-CEDB was proposed for Camellia oleifera fruit detection and reported an mAP of 91.4% [23]. MTD-YOLO was introduced for cherry tomato cluster maturity detection and achieved an mAP of 86.6% [24]. PG-YOLO incorporated ghost modules, backbone simplification, pruning, and knowledge distillation for pomegranate detection, achieving an mAP of 0.934 while reducing model size and improving inference speed [25]. These studies show that lightweight YOLO-based detectors are effective for fruit detection, but their performance may still degrade when targets are small, densely distributed, occluded, or visually similar to the background.

Attention mechanisms, feature-fusion modules, and Transformer-based components have also been explored to enhance agricultural object detection. Spatiotemporal convolutional neural networks combined with Transformers have been applied to pineapple-picking perception, achieving an mAP of 92.54% [26]. Olive-EfficientDet integrated the Convolutional Block Attention Module (CBAM) and a Bidirectional Feature Pyramid Network (Bi-FPN), reporting mAP values above 93% under occlusion and lighting variation [27]. Tracking-based perception has also been investigated for agricultural applications, including DeepSORT-SCAE for young honey peach tracking [28] and three-dimensional relocalization with neural-network-based yield regression for orange detection and tracking [29]. These works suggest that contextual information, multi-scale representation, and attention-based feature refinement are important for detecting fruits in cluttered natural scenes.

Deep learning has also been used for fruit classification and plant disease recognition. Convolutional neural networks have been applied to olive fruit identification under illumination variation [30], potato leaf disease detection using MDSCIRNet [31], and olive cultivar identification with accuracies above 95% [32]. Although these classification-oriented methods demonstrate the effectiveness of deep neural networks in agricultural image analysis, they do not directly address the localization and stage-specific recognition requirements of fruit detection in orchard images. For lychee monitoring, object-level detection is required because the system must localize blossoms and fruit clusters while distinguishing their developmental stages.

The development of intelligent agricultural systems is also closely related to robotic perception and decision-support technologies. Agricultural robots have been reviewed as an important direction for transforming farming practices [5]. YOLOv8-based pest detection has been investigated for precision-agriculture and food-security applications [6]. In addition, smart agro-ecosystems based on sensing, robotic systems, and decision-support models have been discussed for next-generation agriculture [7]. These studies indicate that reliable visual perception is an essential component of practical agricultural automation.

In summary, existing studies have made substantial progress in fruit detection, lightweight YOLO design, attention-based feature enhancement, and agricultural automation. However, lychee growth-stage detection remains less explored than the detection of more common fruits, particularly for a complete stage range from blossom to ripe fruit. This task is challenging because lychee targets vary considerably in scale, color, texture, density, and visibility. Blossom regions are small and sparse, green fruits are visually similar to leaves, young fruits are densely clustered, and ripe fruits may be affected by strong illumination and occlusion. Motivated by these challenges, this study develops and evaluates YOLO-Lychee, a YOLOv8s-based detector that replaces the original SPPF module with SC3T, introduces CAM-based contextual feature aggregation, and adopts EIoU-based bounding-box regression for multi-stage lychee detection under realistic orchard conditions.

3. Materials and Methods

3.1. Lychee Dataset

A stage-specific lychee dataset was constructed for detecting lychee developmental stages under natural orchard conditions. The dataset covers four phenological categories: blossom, young fruit, green fruit, and ripe fruit. These categories represent the principal visual transitions from flowering to harvest maturity and are relevant to stage-specific orchard monitoring, crop management, and harvest planning.

The images were acquired from commercial lychee orchards in Hai Duong and Bac Giang provinces, Vietnam. The collected scenes exhibit substantial variations in canopy density, natural illumination, viewing angle, object scale, target density, partial occlusion, and background appearance. These variations reflect practical challenges encountered in vision-based monitoring of lychee orchards.

The originally acquired dataset comprised 1,145 images and 19,422 manually annotated instances. Before any data augmentation was applied, the images were divided into 802 training images, 115 validation images, and 228 test images, corresponding to approximately 70.0%, 10.0%, and 20.0% of the original dataset, respectively. The partitioning was completed before augmentation to prevent augmented versions of the same image from appearing in different subsets. The validation subset was used for checkpoint selection, whereas the held-out test subset was reserved for final performance evaluation and was not used for model selection or hyperparameter adjustment.

All original images were manually annotated using axis-aligned bounding boxes. Each visually distinguishable target was assigned one of the four developmental-stage labels. When adjacent fruits were sufficiently separable, they were annotated individually. In densely overlapping regions where individual fruits could not be reliably separated, the visible group was annotated as a single target cluster. The same annotation protocol was consistently applied to the training, validation, and test subsets. The annotations were subsequently exported in YOLO-compatible format.

During model training and evaluation, images were processed at an input resolution of 640×640 pixels, and the associated bounding-box coordinates were transformed accordingly. Because blossom was the most under-represented and visually challenging category, exposure-based augmentation was applied exclusively to blossom images in the training subset. This photometric transformation modifies image exposure and brightness without altering target geometry; therefore, the original bounding-box coordinates were retained for the augmented images.

The exposure augmentation generated 179 additional training images containing 839 blossom annotations. Consequently, the training subset increased from 802 to 981 images and from 12,953 to 13,792 annotations. No augmented samples were introduced into the validation or test subsets. The final experimental corpus therefore comprised 1,324 images and 20,261 annotations, of which 1,145 images and 19,422 annotations originated from the acquired orchard dataset.

Representative annotations for three developmental stages are shown in Fig. 1. The young-fruit category was included in all training and evaluation procedures but is omitted from this illustrative figure for compactness.

Table 1 summarizes the composition of the dataset. Panel A reports the class-wise distribution of annotations in the final experimental corpus, including the augmented blossom samples used during training. Panel B distinguishes the originally acquired data from the additional training samples generated by exposure augmentation. In this study, the term “instance” refers to an annotated individual target or a visually inseparable target cluster, rather than an image.

Despite the targeted training augmentation, the final experimental corpus remains class-imbalanced. Blossom accounts for 9.64% of all annotations, whereas ripe fruit represents 44.65%, corresponding

(a) Blossom **(b)** Young fruit **(c)** Green fruit **(d)** Ripe fruit

Figure 1. Representative annotated images from the lychee dataset: (a) blossom, (b) young fruit, (c) green fruit, and (d) ripe fruit. The images illustrate variations in target scale, density, illumination, occlusion, and background appearance under natural orchard conditions

Table 1. Composition of the original lychee dataset and the final experimental corpus. Exposure-based augmentation was applied only to blossom images in the training subset

Panel A. Class-wise annotation distribution in the final experimental corpus

Developmental stage	Train	Validation	Test	Total	Share (%)
Blossom	1,678	84	192	1,954	9.64
Young fruit	3,513	505	897	4,915	24.26
Green fruit	3,402	247	696	4,345	21.45
Ripe fruit	5,199	1,536	2,312	9,047	44.65
Total annotations	**13,792**	**2,372**	**4,097**	**20,261**	**100.00**

Panel B. Accounting of originally acquired and augmented data

Dataset statistic	Train	Validation	Test	Total	Remark
Originally acquired images	802	115	228	1,145	Field data
Exposure-augmented images	179	0	0	179	Training only
Final number of images	**981**	**115**	**228**	**1,324**	Experimental corpus
Annotations in original images	12,953	2,372	4,097	19,422	Manual labels
Additional blossom annotations	839	0	0	839	In augmented images
Final number of annotations	**13,792**	**2,372**	**4,097**	**20,261**	Experimental corpus

to an approximately 4.63-fold difference between the smallest and largest classes. This imbalance reflects both the original orchard sampling conditions and the substantial differences in target density and visibility across developmental stages.

The exposure-based augmentation increases the representation and photometric diversity of blossom samples during training without altering the validation and test distributions. Accordingly, both aggregate and class-wise detection results are reported in the experimental results section to provide a more informative assessment of model performance across the four developmental stages.

3.2. YOLOv8 Model

YOLOv8 is a representative one-stage object detector developed by Ultralytics and has been widely used as a practical baseline for real-time object detection tasks [16, 33]. Compared with earlier YOLO variants, YOLOv8 adopts an efficient convolutional architecture, an anchor-free detection head, and a decoupled prediction structure, providing a favorable balance between detection accuracy and inference speed. These properties make YOLOv8 suitable for agricultural perception tasks, where object detectors must operate under complex backgrounds, scale variation, occlusion, and illumination changes.

As shown in Figure 2, the YOLOv8 architecture consists of three main parts: backbone, neck, and detection

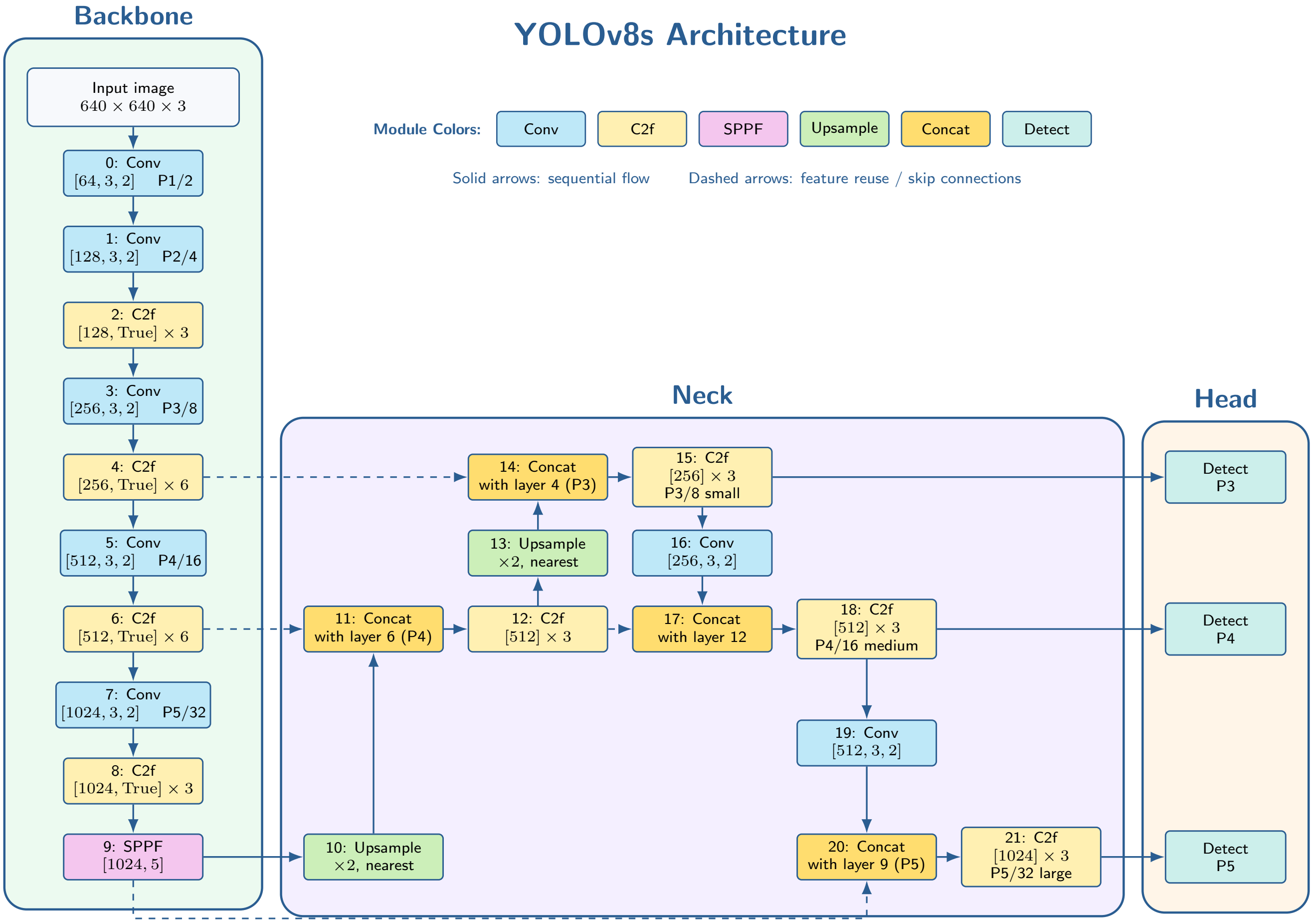

Figure 2. Architecture of the YOLOv8 detector

head. The backbone extracts hierarchical visual features from the input image using convolutional layers and Cross Stage Partial-style feature aggregation. The C2f module is used to improve gradient flow and feature reuse while maintaining computational efficiency. The Spatial Pyramid Pooling-Fast (SPPF) module further enlarges the receptive field and captures multi-scale contextual information without introducing excessive computational overhead.

The neck aggregates feature maps from different stages of the backbone to enhance multi-scale representation. This design is important for detecting objects with large scale variation, such as lychee blossoms, young fruits, green fruits, and ripe fruits. The detection head then predicts object categories and bounding-box locations at multiple feature scales. In contrast to anchor-based YOLO variants, YOLOv8 uses an anchor-free detection strategy, which simplifies the detection pipeline and avoids the need for predefined anchor-box settings.

For bounding-box regression, YOLOv8 employs IoU-based localization objectives together with distribution-based regression components in its training loss. In many YOLO-based implementations, Complete Intersection over Union (CIoU) is commonly used as an IoU-based bounding-box regression loss because it jointly considers the overlap area, center-point distance, and aspect-ratio consistency between the predicted and ground-truth boxes [34]. For a predicted box $B = (x, y, w, h)$ and a ground-truth box $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$, the CIoU loss can be formulated as

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU}\left(B, B^{\mathrm{gt}}\right) + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{\mathrm{gt}}\right)}{c^2} + \alpha v, \qquad (1)$$

where B and B^{gt} denote the predicted and ground-truth bounding boxes, respectively, while $\mathbf{b} = (x, y)$ and $\mathbf{b}^{\mathrm{gt}} = (x^{\mathrm{gt}}, y^{\mathrm{gt}})$ represent their center points. The term $\rho^2\left(\mathbf{b}, \mathbf{b}^{\mathrm{gt}}\right)$ denotes the squared Euclidean distance between the two box centers, and c is the diagonal length of the smallest enclosing box covering both B and

B^{gt}. The aspect-ratio consistency term v is defined as

$$v = \frac{4}{\pi^2}\left(\arctan\left(\frac{w^{gt}}{h^{gt}}\right) - \arctan\left(\frac{w}{h}\right)\right)^2, \quad (2)$$

where w and h denote the width and height of the predicted box, whereas w^{gt} and h^{gt} denote those of the ground-truth box. The trade-off coefficient α, which balances the overlap and aspect-ratio terms, is computed as

$$\alpha = \frac{v}{1 - \mathrm{IoU}(B, B^{gt}) + v}. \quad (3)$$

Although CIoU improves bounding-box regression by considering geometric consistency, its aspect-ratio term may not fully decouple width and height differences. This motivates the use of EIoU in the proposed detector, as described in Section 3.3.

3.3. Proposed YOLO-Lychee Detector

YOLOv8 provides several model scales, including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. In this study, YOLOv8s is selected as the baseline because it provides a practical compromise between detection accuracy and computational cost. This characteristic is important for orchard monitoring and robotic harvesting scenarios, where the detector should remain sufficiently accurate while maintaining real-time inference capability.

The proposed model, named YOLO-Lychee, is designed to improve lychee growth-stage detection under natural orchard conditions. As illustrated in Figure 3, YOLO-Lychee replaces the original SPPF module in the YOLOv8s backbone with SC3T to enhance spatial–channel feature representation. CAM is introduced into the neck to strengthen contextual feature aggregation from multiple receptive fields. In addition, EIoU is adopted for bounding-box regression to improve localization accuracy, especially for small, dense, and partially occluded lychee targets.

The motivation for this design comes from the visual characteristics of lychee orchards. Blossom regions are small and sparse, green fruits are visually similar to leaves, young fruits are densely clustered, and ripe fruits may be affected by strong illumination and partial occlusion. Therefore, the detector requires both local detail preservation and contextual feature modeling. The SC3T module is used to improve feature encoding for small and visually ambiguous targets, while CAM enhances multi-scale contextual representation in the feature-fusion stage. The following subsections describe the main components of the proposed detector.

SC3T Module. The Spatial and Channel Cross-Transformer (SC3T) module was originally introduced in our previous work for improving the detection of small and distant objects in UAV-based imagery [35]. In this study, SC3T is incorporated into the YOLOv8s backbone to enhance feature representation for lychee targets with small size, dense distribution, and partial occlusion. The module combines the multi-scale receptive-field capability of Spatial Pyramid Pooling (SPP) [36] with the contextual modeling ability of a Transformer-based C3TR structure.

As shown in Figure 4, the SPP component extracts multi-scale spatial information by applying pooling operations with different receptive fields. This design helps the network capture both fine local details and broader contextual cues, which are useful for detecting lychee objects with large scale variation. For example, blossom regions usually occupy small image areas, whereas fruit clusters may appear at larger scales and with different densities. Multi-scale pooling therefore improves the ability of the backbone to represent objects at different developmental stages.

The C3TR component introduces Transformer-based feature modeling into the convolutional feature-extraction pipeline. By using self-attention, the module can capture long-range dependencies and strengthen relationships among spatially separated regions in the feature map. This is beneficial in orchard images, where lychee fruits may be partially occluded by leaves or distributed across dense clusters. Compared with purely convolutional feature extraction, the Transformer-based branch provides additional contextual information that can help distinguish lychee targets from visually similar backgrounds.

By combining SPP and C3TR, the SC3T module enhances both multi-scale spatial representation and contextual feature interaction. In the proposed YOLO-Lychee detector, SC3T is used to strengthen backbone features before multi-scale feature fusion in the neck. Its contribution to detection performance is quantitatively evaluated in the ablation study in Section 4.3.

CAM Module. The Context Augmentation Module (CAM) was introduced in the Complementary Feature Pyramid Network (CFPN) to improve tiny-object detection by enriching contextual information from multiple receptive fields [37]. In lychee orchard images, small blossoms and young fruits are often surrounded by leaves, branches, and visually similar backgrounds. Therefore, contextual feature aggregation is important for improving the discriminability of small and partially occluded lychee targets. In the proposed YOLO-Lychee detector, CAM is incorporated into the neck to enhance multi-scale feature fusion before prediction.

As shown in Figure 5, CAM adopts a multi-branch convolutional structure to extract contextual features at different receptive-field scales. Each branch

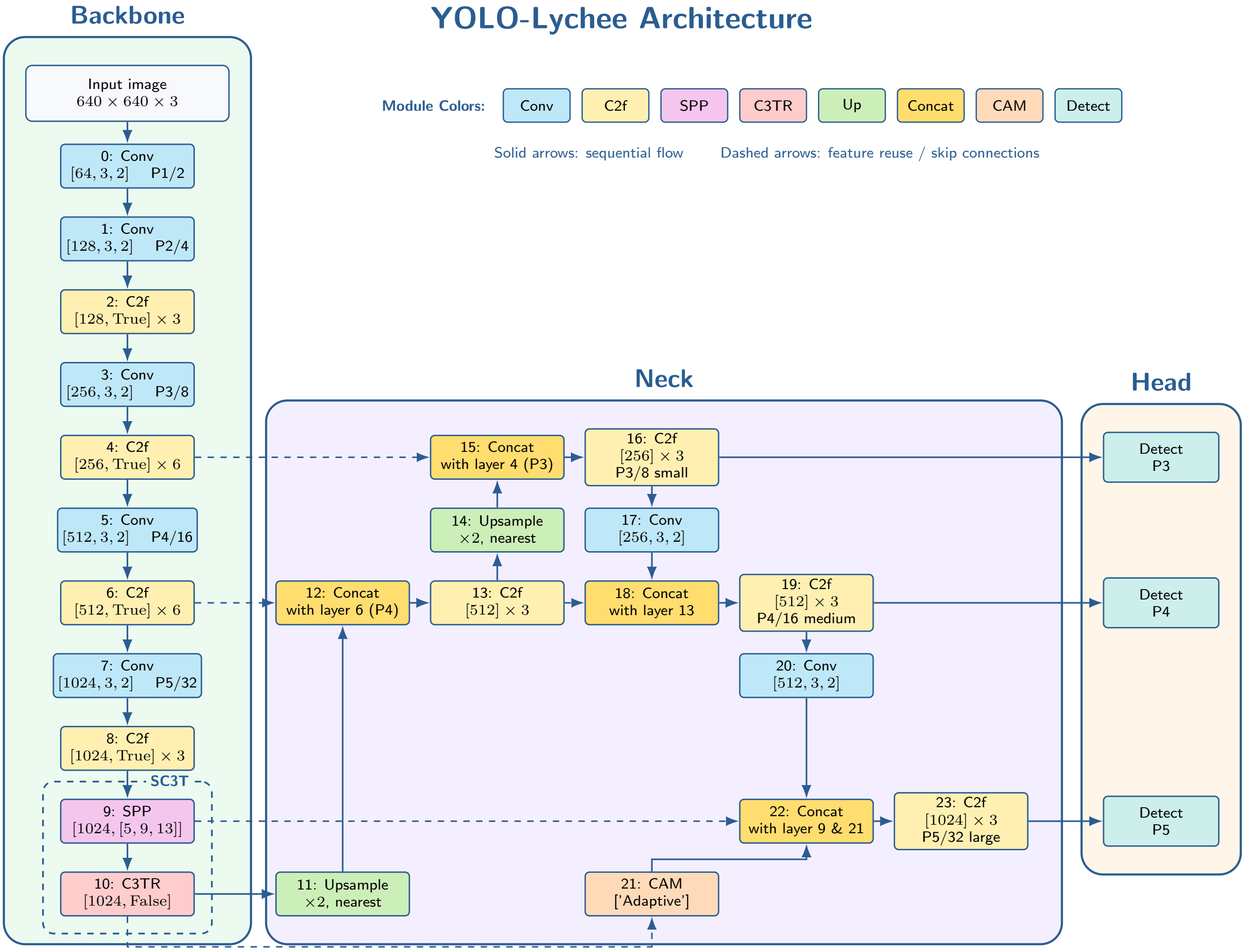

Figure 3. Architecture of the proposed YOLO-Lychee detector. The model is built on the YOLOv8s scale and integrates SC3T in the backbone for spatial–channel feature enhancement, CAM in the neck for contextual feature aggregation, and EIoU as the bounding-box regression loss

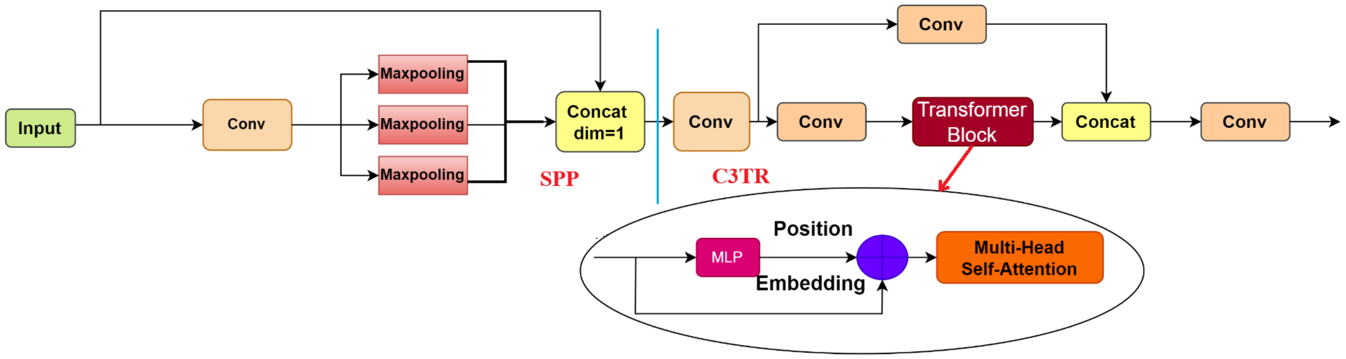

Figure 4. Architecture of the SC3T module, which combines spatial pyramid pooling and Transformer-based feature modeling for enhanced spatial–channel representation

applies a 3×3 convolution with a different dilation rate, such as 1, 3, and 5. The branch with a small dilation rate captures local details, while branches with larger dilation rates provide broader contextual

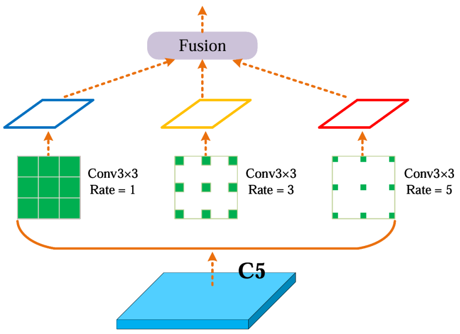

Figure 5. Architecture of the CAM module [37]

information from the surrounding regions. This design enables the detector to jointly model object-level details and neighboring context, which is beneficial for distinguishing lychee targets from cluttered orchard backgrounds.

The features extracted from different branches are then fused to form an enhanced contextual representation. CAM supports several fusion strategies, including concatenation fusion, adaptive fusion, and weighted fusion. Concatenation fusion directly combines branch features along the channel dimension, increasing the feature capacity but also increasing the number of parameters. Adaptive fusion learns branch-specific weights through an auxiliary sub-network, allowing the model to adjust the contribution of each receptive field according to the input features. Weighted fusion computes spatial or channel weights and combines the branch features according to their relative importance.

In this study, different CAM fusion strategies are evaluated in the ablation study to analyze their effects on lychee detection performance. The experimental results show that adaptive fusion provides the best overall balance between detection accuracy and model complexity. Therefore, the adaptive CAM configuration is used in the final YOLO-Lychee detector. By introducing CAM into the neck, the proposed model strengthens contextual feature aggregation and improves the detection of small, dense, and visually ambiguous lychee targets under natural orchard conditions.

EIoU Loss Function. Accurate bounding-box regression is important for lychee growth-stage detection because lychee targets often appear as small, dense, and partially occluded objects in natural orchard scenes. YOLOv8 commonly adopts IoU-based regression objectives, in which the localization loss is designed to maximize the spatial agreement between a predicted bounding box and its corresponding ground-truth box. In this study, the Efficient Intersection over Union (EIoU) loss [38] is adopted to improve localization accuracy by explicitly decomposing the regression error into three geometric components: overlap area, center-point distance, and side-length discrepancy.

Let $B = (x, y, w, h)$ denote a predicted bounding box and $B^{\mathrm{gt}} = (x^{\mathrm{gt}}, y^{\mathrm{gt}}, w^{\mathrm{gt}}, h^{\mathrm{gt}})$ denote the corresponding ground-truth box, where (x, y) and $(x^{\mathrm{gt}}, y^{\mathrm{gt}})$ are the respective center coordinates, and w, h, w^{gt}, and h^{gt} are the corresponding box widths and heights. Let $\mathbf{b} = (x, y)$ and $\mathbf{b}^{\mathrm{gt}} = (x^{\mathrm{gt}}, y^{\mathrm{gt}})$ represent the center points of B and B^{gt}, respectively. The smallest enclosing box that covers both B and B^{gt} is denoted by C, with width w_c and height h_c. The EIoU loss is formulated as

$$L_{\mathrm{EIoU}} = L_{\mathrm{IoU}} + L_{\mathrm{dis}} + L_{\mathrm{side}}, \tag{4}$$

where L_{IoU} penalizes insufficient overlap, L_{dis} penalizes the discrepancy between the box centers, and L_{side} penalizes differences in the widths and heights of the predicted and ground-truth boxes.

The overlap loss is defined as

$$L_{\mathrm{IoU}} = 1 - \mathrm{IoU}\left(B, B^{\mathrm{gt}}\right) = 1 - \frac{\left|B \cap B^{\mathrm{gt}}\right|}{\left|B \cup B^{\mathrm{gt}}\right|}, \tag{5}$$

where $\left|B \cap B^{\mathrm{gt}}\right|$ and $\left|B \cup B^{\mathrm{gt}}\right|$ denote the intersection and union areas of the predicted and ground-truth boxes, respectively. A smaller value of L_{IoU} therefore indicates a greater spatial overlap between the two boxes.

The normalized center-distance loss is expressed as

$$L_{\mathrm{dis}} = \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{\mathrm{gt}}\right)}{w_c^2 + h_c^2}, \tag{6}$$

where

$$\rho^2\left(\mathbf{b}, \mathbf{b}^{\mathrm{gt}}\right) = \left(x - x^{\mathrm{gt}}\right)^2 + \left(y - y^{\mathrm{gt}}\right)^2 \tag{7}$$

is the squared Euclidean distance between the predicted and ground-truth box centers. The denominator $w_c^2 + h_c^2$ corresponds to the squared diagonal length of the smallest enclosing box C, thereby normalizing the center-distance penalty with respect to object scale.

The side-length loss is defined as

$$L_{\mathrm{side}} = \frac{\left(w - w^{\mathrm{gt}}\right)^2}{w_c^2} + \frac{\left(h - h^{\mathrm{gt}}\right)^2}{h_c^2}. \tag{8}$$

This term directly minimizes the discrepancies between the predicted and ground-truth box dimensions, allowing the width and height errors to be optimized independently.

This term directly penalizes the width and height discrepancies between the predicted and ground-truth boxes. Unlike CIoU [34], which uses an aspect-ratio consistency term, EIoU separately optimizes the width and height errors. This decoupled formulation is beneficial for detecting lychees at different developmental stages, where object shapes and scales vary substantially from blossom clusters to young, green, and ripe fruits.

By combining overlap, center-distance, and side-length penalties, EIoU provides a more explicit regression objective for bounding-box localization. In the proposed YOLO-Lychee detector, this loss is used to improve the localization of small and partially occluded lychee targets while keeping the model architecture unchanged. The contribution of EIoU is further examined in the ablation study, where it is compared with other IoU-based losses, including GIoU [39], WIoU [40], CIoU [34], and SIoU [41].

3.4. Evaluation Metrics

The detection performance of the evaluated models was measured using standard object-detection metrics, including Precision, Recall, F1-score, mAP_{50}, and $\mathrm{mAP}_{50:95}$. In addition, model efficiency was assessed using the number of parameters, model size, floating-point operations (FLOPs), inference time, and frames per second (FPS). These metrics provide a comprehensive evaluation of both detection accuracy and deployment feasibility.

Intersection over Union (IoU) quantifies the spatial overlap between a predicted bounding box B and its corresponding ground-truth box B^{gt}. It is defined as

$$\mathrm{IoU}\left(B, B^{\mathrm{gt}}\right) = \frac{\left|B \cap B^{\mathrm{gt}}\right|}{\left|B \cup B^{\mathrm{gt}}\right|}, \tag{9}$$

where $\left|B \cap B^{\mathrm{gt}}\right|$ and $\left|B \cup B^{\mathrm{gt}}\right|$ denote the intersection and union areas of the two boxes, respectively. A predicted bounding box is classified as a true positive when it is assigned to the correct class and its IoU with a matched ground-truth box is greater than or equal to a predefined threshold. Otherwise, it is treated as a false positive, while an unmatched ground-truth object is counted as a false negative.

Precision and recall are computed as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{10}$$

and

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{11}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. Precision represents the proportion of correct detections among all predicted objects, whereas recall represents the proportion of ground-truth objects that are successfully detected.

The F1-score is the harmonic mean of precision and recall and is used to evaluate the balance between these two metrics. It is defined as

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{12}$$

Average Precision (AP) summarizes the precision–recall curve for each object class at a specified IoU threshold. For class i and IoU threshold τ, AP can be approximated using a discretized precision–recall curve as

$$\mathrm{AP}_i(\tau) = \sum_n (R_n - R_{n-1})\, P_{\mathrm{interp},n}, \tag{13}$$

where R_n denotes the recall at the n-th operating point and $P_{\mathrm{interp},n}$ denotes the corresponding interpolated precision. The class index is denoted by i, while τ specifies the IoU threshold used to determine whether a detection is considered correct.

The mean Average Precision at an IoU threshold of 0.50, denoted by mAP_{50}, is calculated as

$$\mathrm{mAP}_{50} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i(0.50), \tag{14}$$

where N denotes the number of object classes. In this study, $N = 4$, corresponding to the blossom, young-fruit, green-fruit, and ripe-fruit classes.

A more stringent metric, $\mathrm{mAP}_{50:95}$, averages the AP values over IoU thresholds ranging from 0.50 to 0.95 in increments of 0.05. It is defined as

$$\mathrm{mAP}_{50:95} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \left[\frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i(\tau) \right], \tag{15}$$

where $\mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}$ is the set of evaluated IoU thresholds and $|\mathcal{T}| = 10$. Compared with mAP_{50}, $\mathrm{mAP}_{50:95}$ provides a stricter assessment of localization performance because detections must maintain accurate bounding-box alignment over progressively higher IoU thresholds.

Model efficiency was evaluated using the number of parameters and checkpoint size as measures of model compactness, together with floating-point operations (FLOPs) as an estimate of computational complexity. The average inference time was measured in milliseconds per image, and the corresponding inference throughput was calculated as

$$\mathrm{FPS} = \frac{1000}{t_{\mathrm{inf}}}, \tag{16}$$

where t_{inf} denotes the average inference time per image in milliseconds. All models were evaluated under the same hardware, software, input-resolution, batch-size, and measurement settings to ensure a fair comparison. These efficiency measures were analyzed together with the detection metrics to assess the trade-off between predictive accuracy and computational cost.

4. Experimental Results

4.1. Experimental Setup

All experiments were implemented using the Ultralytics YOLO framework with PyTorch and CUDA acceleration. Both model training and testing were conducted on an NVIDIA A100 GPU. The proposed detector was developed based on the YOLOv8s scale and trained on the lychee dataset described in Section 3.1. Four developmental stages were considered: blossom, young fruit, green fruit, and ripe fruit. To ensure a fair comparison, all model variants were trained and evaluated using the same hardware environment, data split, input resolution, optimizer, and data-augmentation protocol.

Input images were resized to 640×640 pixels. Each model was trained for 100 epochs with a batch size of 16 using stochastic gradient descent (SGD). The initial learning rate, momentum, and weight decay were set to 0.01, 0.937, and 0.0005, respectively. Standard training augmentations, including HSV perturbation, random translation, random scaling, horizontal flipping, and mosaic augmentation, were applied to improve robustness to illumination changes, scale variations, and complex backgrounds in natural orchard scenes. Mosaic augmentation was disabled during the final 10 epochs to stabilize late-stage optimization.

For the loss-function analysis, CIoU, GIoU, EIoU, and SIoU were evaluated under the same training protocol. The best checkpoint of each model was selected according to its validation performance and subsequently evaluated on the independent test set using the NVIDIA A100 GPU. Detection performance was reported in terms of Precision, Recall, mAP_{50}, and $mAP_{50:95}$. Model efficiency was further evaluated using the number of parameters, computational complexity, checkpoint size, inference latency, and frames per second (FPS). Inference latency and FPS were measured on the same NVIDIA A100 GPU to ensure consistent efficiency comparisons among the evaluated models. The principal experimental settings are summarized in Table 2.

4.2. Comparison with Recent YOLO Detectors

To provide a comprehensive benchmark against recent YOLO-family detectors, YOLO-Lychee was compared with YOLOv8s [16], YOLOv9s [17], YOLOv10s [18], YOLO11s [19], YOLOv12s [20], and YOLO26s [21]. All detectors were trained and evaluated on the same lychee dataset using the experimental protocol described in Section 4.1. The comparison considers both detection accuracy and computational efficiency, including the number of parameters, model size, FLOPs, FPS, Precision, Recall, mAP_{50}, and $mAP_{50:95}$. For YOLO-Lychee, the detection metrics are reported as mean ± standard deviation over six random seeds, namely 0, 1, 2, 3, 4, and 5, to assess the stability of the proposed detector under different initialization settings.

As shown in Table 3, YOLO-Lychee achieved the highest mAP_{50} among all compared detectors, with a mean value of $88.42 \pm 0.38\%$. Compared with YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, YOLOv12s, and YOLO26s, the proposed detector improved mAP_{50} by 1.74, 0.62, 1.85, 1.05, 0.21, and 2.86 percentage points, respectively. The relatively small standard deviation over six random seeds indicates that the proposed configuration provides stable performance at the standard IoU threshold. This improvement suggests that the integration of SC3T, CAM, and EIoU is beneficial for detecting lychee objects across different growth stages in complex orchard scenes.

For the stricter $mAP_{50:95}$ metric, YOLO-Lychee achieved $72.76 \pm 0.45\%$, which is competitive with recent YOLO-family detectors. It outperformed YOLOv8s, YOLOv9s, YOLOv10s, YOLOv12s, and YOLO26s by 0.95, 0.08, 1.47, 0.06, and 1.26 percentage points, respectively. The best $mAP_{50:95}$ was obtained by YOLO11s, which exceeded YOLO-Lychee by only 0.15 percentage points. These results indicate that the proposed detector provides competitive localization performance under stricter IoU thresholds. Nevertheless, the small differences among YOLOv9s, YOLO11s, YOLOv12s, and YOLO-Lychee also suggest that further localization refinement is needed to obtain a more decisive advantage in $mAP_{50:95}$.

In terms of Precision and Recall, YOLO-Lychee achieved $83.02 \pm 2.38\%$ and $82.19 \pm 1.87\%$, respectively. Its Precision was lower than that of most compared detectors, except YOLOv12s, whereas its Recall was higher than those of YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, and YOLO26s, but lower than that of YOLOv12s. This behavior indicates that the proposed detector tends to improve mAP-based detection performance, particularly mAP_{50}, while maintaining a reasonable balance between missed detections and false positives. The results also show that high mAP does not necessarily require the highest Precision or Recall individually, since mAP summarizes the detector's ranking quality across confidence thresholds.

Regarding computational efficiency, YOLO-Lychee required 20.26M parameters, 38.90 MB of model storage, and 28.40G FLOPs, which are higher than those of the compared YOLO-family baselines. This increase is mainly due to the additional SC3T and CAM components introduced to enhance contextual feature representation and small-object detection. Despite the increased complexity, YOLO-Lychee still achieved 230.3 ± 17.47 FPS on the tested GPU, indicating that it remains capable of real-time inference. Among the compared models, YOLOv9s had the smallest number of parameters and model size, YOLOv12s required the lowest FLOPs, YOLOv10s achieved the highest FPS,

Table 2. Main experimental settings

Item	Setting
Framework	Ultralytics YOLO with PyTorch and CUDA
Training and testing hardware	NVIDIA A100 GPU
Model	YOLO-Lychee
Backbone scale	YOLOv8s
Input resolution	640 × 640
Number of classes	4
Epochs	100
Batch size	16
Optimizer	SGD
Initial learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Main augmentations	HSV, translation, scaling, horizontal flip, mosaic
Mosaic deactivation	Final 10 epochs
Checkpoint selection	Best validation performance
Evaluation set	Independent test set

Table 3. Comparison with recent YOLO-family detectors on the lychee test set. Best values are shown in bold. The symbols ↑ and ↓ indicate whether higher or lower values are preferred, respectively. For YOLO-Lychee, detection metrics are reported as mean ± standard deviation over six random seeds

Model	Params (M) ↓	Size (MB) ↓	FLOPs (G) ↓	FPS (↑)	P (%) ↑	R (%) ↑	mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑
YOLOv8s [16]	11.14	21.48	14.33	302.6 ± 8.54	85.46	81.44	86.68	71.81
YOLOv9s [17]	**7.29**	**14.53**	13.70	163.8 ± 2.41	85.81	81.42	87.80	72.68
YOLOv10s [18]	8.07	15.77	12.39	**302.7** ± 3.66	84.29	79.80	86.57	71.29
YOLO11s [19]	9.43	18.30	10.78	273.3 ± 1.58	**86.63**	79.76	87.37	**72.91**
YOLOv12s [20]	9.25	18.06	**10.76**	201.2 ± 2.07	82.96	**83.64**	88.21	72.70
YOLO26s [21]	9.95	19.38	11.26	255.7 ± 7.07	84.20	80.62	85.56	71.50
YOLO-Lychee (Ours)	20.26	38.90	28.40	230.3 ± 17.47	83.02 ± 2.38	82.19 ± 1.87	**88.42** ± 0.38	72.76 ± 0.45

and YOLO11s obtained the highest $mAP_{50:95}$. Overall, YOLO-Lychee provides the best mAP_{50} and competitive $mAP_{50:95}$, while its higher computational cost reflects a trade-off between detection accuracy and model compactness.

To further examine the behavior of the proposed detector across individual lychee growth stages, Table 4 reports the class-wise performance of YOLO-Lychee on the test set. This analysis is important because the four target classes differ substantially in object scale, visual appearance, density, and similarity to the orchard background. The overall row summarizes the aggregate performance on the full test set, whereas the remaining rows show the detection performance for each developmental stage.

As shown in Table 4, YOLO-Lychee achieved strong performance on the three fruit-stage categories. Green fruit obtained the highest Recall, mAP_{50}, and $mAP_{50:95}$, reaching 93.40%, 96.20%, and 83.00%, respectively. This result suggests that the proposed detector can effectively recognize green lychee clusters, even though their color and texture may be visually similar to surrounding foliage in natural orchard scenes.

The ripe and young fruit classes also achieved stable detection performance. Ripe fruit obtained an mAP_{50} of 92.90% and an $mAP_{50:95}$ of 78.50%, while young fruit achieved an mAP_{50} of 92.58% and an $mAP_{50:95}$ of 78.54%. These results indicate that the proposed feature-enhancement strategy can capture discriminative visual cues for fruit-stage lychees under variations in scale, density, and partial occlusion.

In contrast, blossom remained the most challenging class. Although the blossom class achieved an mAP_{50} of 72.00%, its Recall and $mAP_{50:95}$ were substantially lower than those of the fruit-stage classes. This performance gap can be attributed to the small size, sparse distribution, irregular shape, and strong background interference of blossom regions. In addition, blossom clusters often contain fine-grained textures that are more sensitive to illumination changes, making accurate localization difficult under complex orchard conditions.

Overall, the class-wise results demonstrate that YOLO-Lychee performs reliably for green, ripe, and young lychee detection, while blossom detection remains the main source of performance degradation. This finding is consistent with the class imbalance

Table 4. Class-wise performance of YOLO-Lychee on the lychee test set. The symbol ↑ indicates that higher values are preferred

Class	Instances	P (%) ↑	R (%) ↑	mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑
Blossom	192	76.80	62.00	72.00	51.00
Green fruit	696	86.00	93.40	96.20	83.00
Ripe fruit	2,312	84.00	87.50	92.90	78.50
Young fruit	897	85.28	85.86	92.58	78.54
Overall	4,097	83.02	82.19	88.42	72.76

and visual complexity of the dataset. Future work should therefore focus on improving early-stage lychee recognition through small-object feature refinement, imbalance-aware training strategies, and more robust localization mechanisms for visually ambiguous blossom regions.

4.3. Ablation Study

To verify the contribution of each component in YOLO-Lychee, we conducted a component-wise ablation study using YOLOv8s as the baseline. The evaluated components include the SC3T module, the CAM module, and the EIoU bounding-box regression loss. Since the original YOLOv8s uses CIoU as the default localization loss, the ablation was designed progressively. First, SC3T was added to the baseline to enhance spatial–channel feature representation. Second, CAM was incorporated to improve contextual feature aggregation. Finally, CIoU was replaced with EIoU to examine the effect of the localization objective. All variants were trained and evaluated using the same dataset split, input resolution, optimizer, augmentation strategy, and training protocol described in Section 4.1. The results are reported in Table 5.

As shown in Table 5, adding SC3T to the YOLOv8s baseline increased mAP_{50} from 86.68% to 87.19% and $mAP_{50:95}$ from 71.81% to 71.92%. These gains of 0.51 and 0.11 percentage points, respectively, indicate that SC3T provides a positive but moderate improvement in mAP-based detection performance. However, Precision decreased from 85.46% to 82.68%, and Recall decreased from 81.44% to 80.47%. This suggests that SC3T strengthens feature representation but, when used alone, does not fully address the detection challenges caused by small targets, occlusion, and background similarity in natural orchard scenes.

When CAM was further added to the SC3T-enhanced model while retaining CIoU, the detector achieved a Precision of 82.73%, Recall of 81.63%, mAP_{50} of 87.55%, and $mAP_{50:95}$ of 72.05%. Compared with YOLOv8s + SC3T, CAM improved Recall by 1.16 percentage points, mAP_{50} by 0.36 percentage points, and $mAP_{50:95}$ by 0.13 percentage points. These results indicate that contextual feature aggregation is beneficial for retrieving more lychee instances and slightly improving localization quality. The improvement is consistent with the role of CAM in integrating information from multiple receptive fields, which is useful for detecting lychee targets with varying scales, densities, and degrees of occlusion.

After replacing CIoU with EIoU, the full YOLO-Lychee model achieved the best overall performance, with a Recall of 82.19%, mAP_{50} of 88.42%, and $mAP_{50:95}$ of 72.76%. Compared with the SC3T + CAM variant using CIoU, EIoU improved Precision by 0.29 percentage points, Recall by 0.56 percentage points, mAP_{50} by 0.87 percentage points, and $mAP_{50:95}$ by 0.71 percentage points. This confirms that EIoU provides a more effective bounding-box regression objective for lychee growth-stage detection, particularly under stricter IoU thresholds. The improvement in $mAP_{50:95}$ is especially important because it reflects better localization consistency rather than only successful detection at a loose IoU threshold.

Overall, the ablation results show that the three components contribute to the final detector in a complementary manner. SC3T improves spatial–channel feature representation, CAM enhances contextual feature aggregation, and EIoU improves bounding-box localization. Although the baseline YOLOv8s achieved the highest Precision, YOLO-Lychee obtained the best Recall, mAP_{50}, and $mAP_{50:95}$, indicating a more favorable balance between object retrieval and localization accuracy. Therefore, the combination of SC3T, CAM, and EIoU was selected as the final configuration for lychee growth-stage detection.

Effect of Bounding-Box Regression Loss. To further analyze the influence of bounding-box regression objectives, we evaluated five IoU-based losses, including EIoU [38], GIoU [39], WIoU [40], CIoU [34], and SIoU [41]. In this experiment, the detector architecture was fixed as YOLOv8s-SC3T-CAM, and only the localization loss was changed. All other training settings were kept identical to those described in Section 4.1. The results are summarized in Table 6.

As shown in Table 6, EIoU achieved the best mAP-based performance, with an mAP_{50} of 88.42% and an $mAP_{50:95}$ of 72.76%. Compared with CIoU,

Table 5. Component-wise ablation study of YOLO-Lychee on the lychee test set. CIoU is the default bounding-box regression loss in YOLOv8s. Best values are shown in bold. The symbol ↑ indicates that higher values are preferred

Model	SC3T	CAM	CIoU	EIoU	P (%) ↑	R (%) ↑	mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑
YOLOv8s	–	–	√	–	**85.46**	81.44	86.68	71.81
YOLOv8s + SC3T	√	–	√	–	82.68	80.47	87.19	71.92
YOLOv8s + SC3T + CAM	√	√	√	–	82.73	81.63	87.55	72.05
YOLO-Lychee	√	√	–	√	83.02	**82.19**	**88.42**	**72.76**

Table 6. Ablation study of different bounding-box regression losses on the proposed detector. Best values are shown in bold. The symbol ↑ indicates that higher values are preferred

Loss	P (%) ↑	R (%) ↑	mAP_{50} (%) ↑	$mAP_{50:95}$ (%) ↑
EIoU [38]	83.02	82.19	**88.42**	**72.76**
GIoU [39]	**84.66**	80.76	88.40	71.85
WIoU [40]	80.10	80.44	86.60	71.00
CIoU [34]	82.73	81.63	87.55	72.05
SIoU [41]	82.93	**82.39**	87.45	72.30

EIoU improved mAP_{50} by 0.87 percentage points and $mAP_{50:95}$ by 0.71 percentage points. These results indicate that explicitly penalizing center-distance and side-length discrepancies is beneficial for localizing lychee targets, especially when evaluation is performed across multiple IoU thresholds.

GIoU obtained the highest Precision of 84.66% and a competitive mAP_{50} of 88.40%, only 0.02 percentage points lower than EIoU. However, its Recall and $mAP_{50:95}$ were lower than those of EIoU, suggesting that GIoU produces more conservative detections but does not achieve the best localization consistency under stricter thresholds. In contrast, SIoU achieved the highest Recall of 82.39% and a competitive $mAP_{50:95}$ of 72.30%, but its mAP_{50} was lower than that of EIoU by 0.97 percentage points.

WIoU produced the lowest performance among the compared losses, with a Precision of 80.10%, Recall of 80.44%, mAP_{50} of 86.60%, and $mAP_{50:95}$ of 71.00%. Although WIoU introduces a dynamic focusing mechanism for bounding-box regression, it appears less suitable for this dataset, where lychee targets vary substantially in scale, density, occlusion, and visual ambiguity. Overall, EIoU was selected as the final bounding-box regression loss because it achieved the highest mAP_{50} and $mAP_{50:95}$, providing the most effective localization performance for the proposed detector.

4.4. Qualitative Results

To complement the quantitative evaluation, qualitative comparisons were conducted on representative test images from the four lychee developmental stages: blossom, green fruit, ripe fruit, and young fruit. The same input image of each class was used for all detectors to enable a direct visual comparison. Figure 6 shows the detection results of the proposed YOLO-Lychee model and recent YOLO-family detectors, including YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, YOLOv12s, and YOLO26s.

As illustrated in Figure 6, all compared YOLO-family detectors can detect lychee targets across the four developmental stages, but their visual behavior differs under dense clusters, background similarity, and small-object conditions. The proposed YOLO-Lychee detector produces more complete detections in the blossom image, where the targets are small, irregularly shaped, and partially mixed with leaves and branches. Compared with the baseline YOLO variants, the proposed detector identifies a larger number of blossom regions and maintains relatively stable confidence scores for both central and peripheral blossom clusters. This observation supports the role of SC3T and CAM in enhancing spatial–channel representation and contextual feature aggregation for small and visually ambiguous targets.

For green fruit, most detectors localize the main lychee clusters because the fruits are larger and more densely distributed than blossoms. Nevertheless, the proposed detector provides more consistent coverage of adjacent fruits and fruit clusters, including those located near leaf boundaries. This suggests that the additional contextual modeling helps distinguish green lychees from visually similar foliage. Several baseline models also detect many green fruits, but their bounding boxes are less uniform in dense regions, indicating that background similarity and overlapping targets remain challenging.

For ripe lychees, the compared detectors generally identify the prominent fruit clusters in the foreground and middle regions. The proposed detector provides stable detections across fruits with different scales and illumination levels, including partially occluded ripe fruits. However, some small or distant ripe fruits remain difficult for all models, especially in regions affected by strong outdoor lighting or heavy leaf occlusion. This indicates that the proposed model

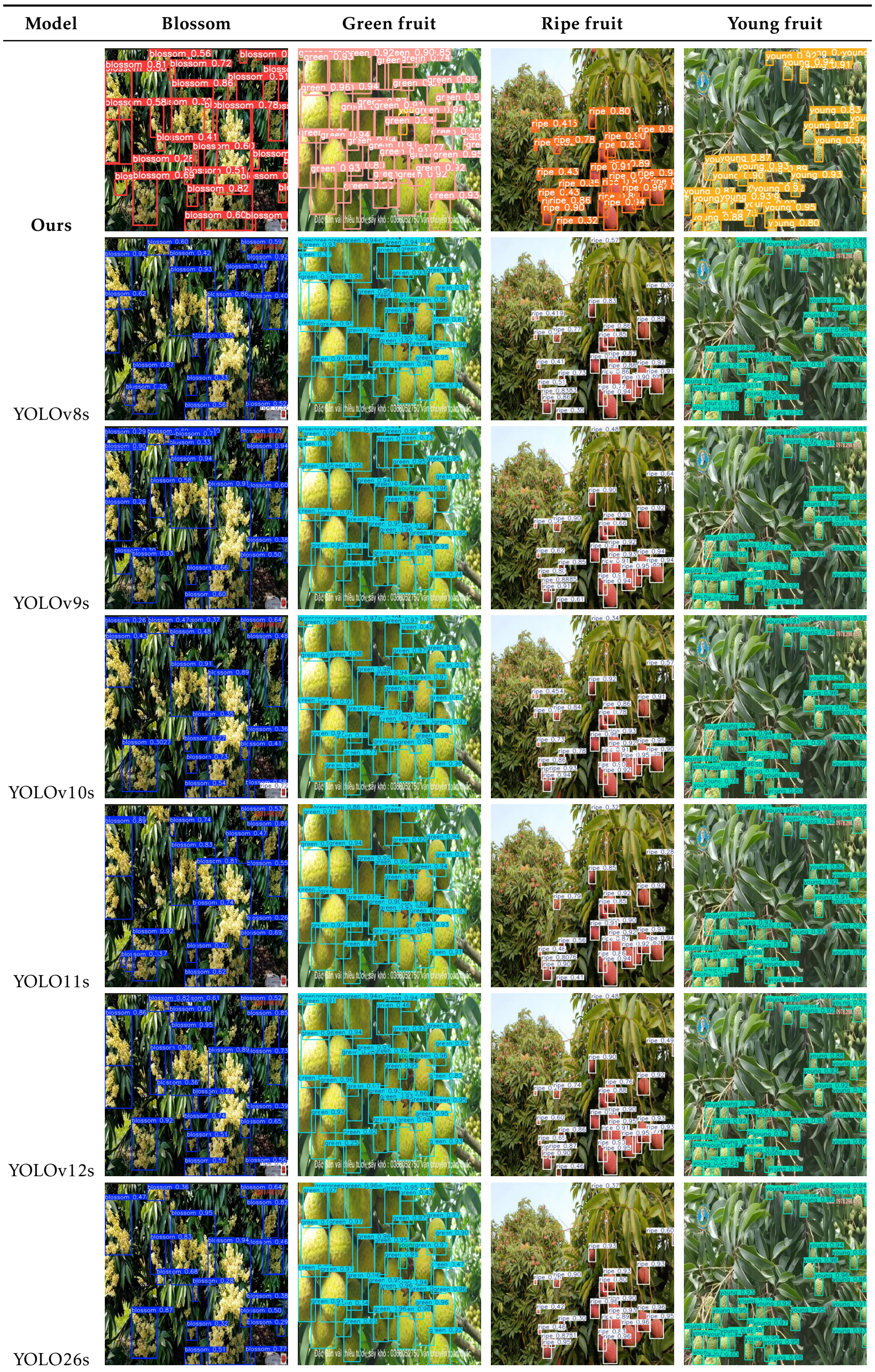

Figure 6. Qualitative comparison of the proposed detector and recent YOLO-family models on representative lychee test images. Each column corresponds to the same input image for all models, enabling direct visual comparison across blossom, green fruit, ripe fruit, and young fruit stages

improves detection completeness, but the localization of small distant fruits is still limited by object scale and scene complexity.

For young lychees, the proposed detector detects multiple fruit clusters with compact bounding boxes and high confidence values. The baseline models also show strong detection ability for this category, reflecting the dense distribution and repeated visual patterns of young fruits. However, some detections are concentrated in cluttered leaf regions, where young fruits and surrounding foliage share similar color and texture. This confirms that young-fruit detection remains sensitive to background interference even when the overall quantitative performance is high.

Overall, the qualitative comparison is consistent with the quantitative results. The proposed YOLO-Lychee detector provides visually stable detections across the four lychee growth stages and shows clear advantages in small-object and dense-cluster scenarios, especially for blossom and clustered green-fruit detection. At the same time, the remaining missed detections and lower-confidence boxes in blossom and distant-fruit regions indicate that further improvements should focus on blossom-specific data expansion, small-object feature preservation, and robustness under strong illumination and occlusion.

5. Conclusion

This study developed YOLO-Lychee for detecting lychee growth stages under natural orchard conditions. By integrating SC3T, CAM, and EIoU into the YOLOv8s framework, the proposed detector improved mAP_{50} and achieved competitive $mAP_{50:95}$ compared with recent YOLO-family models under the same experimental protocol. The ablation results confirmed that spatial–channel modeling, contextual feature aggregation, and improved bounding-box regression contribute complementary benefits to the final model.

The experimental results indicate that YOLO-Lychee is effective for detecting green, ripe, and young lychee fruits in complex orchard scenes involving dense clusters, partial occlusion, and visually similar backgrounds. However, blossom detection remains the most challenging case due to its small scale, sparse appearance, irregular texture, and strong interference from leaves and branches. In addition, the improved accuracy is obtained at the cost of increased model complexity compared with lightweight YOLO baselines.

Future work will focus on improving blossom recognition and small-object localization, strengthening imbalance-aware training, and reducing computational cost through pruning, quantization, or lightweight module design. These improvements will further support practical deployment in orchard monitoring and robotic harvesting systems.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval and consent for participate Not applicable.

Acknowledgments

This research is funded by International School, Vietnam National University, Hanoi, Vietnam.

References

[1] NV Dung, DQ Nghi, NQ Hung, and NQ Huy. Current production and development trends of lychee (litchi chinensis sonn.) in vietnam in the coming future. In *VI International Symposium on Lychee, Longan and Other Sapindaceae Fruits*, pages 7–14, 2019.

[2] Naoshi Kondo, Mitsuji Monta, and Noboru Noguchi. *Agricultural robots: mechanisms and practice*. Apollo Books, 2011.

[3] Inkyu Sa, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool. Deepfruits: A fruit detection system using deep neural networks. *sensors*, 16(8):1222, 2016.

[4] Aleana Gongal, Suraj Amatya, Manoj Karkee, Qin Zhang, and Karen Lewis. Sensors and systems for fruit detection and localization: A review. *Computers and Electronics in Agriculture*, 116:8–19, 2015.

[5] Ata Jahangir Moshayedi, Amir Sohail Khan, Yiguo Yang, Jiandong Hu, and Amin Kolahdooz. Robots in agriculture: Revolutionizing farming practices. *EAI Endorsed Transactions on AI and Robotics*, 3, 2024.

[6] Ameer Tamoor Khan, Sign Marie Jensen, and Noman Khan. Advancing food security through precision agriculture: Yolov8's role in efficient pest detection and management. *EAI Endorsed Transactions on AI and Robotics*, 4, 2025.

[7] Ata Jahangir Moshayedi, Amir Sohail Khan, Amin Kolahdooz, Aiman Elragig, Zeashan Khan, and David Bassir. Smart agro-ecosystem: A review of llm-based robotic systems for sensing and decision support in precision agriculture. *EAI Endorsed Transactions on AI and Robotics*, 5, 2026.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[9] Subramanian Parvathi and Sankar Tamil Selvi. Detection of maturity stages of coconuts in complex background using faster r-cnn model. *biosystems engineering*, 202:119–132, 2021.

[10] Suchet Bargoti and James Underwood. Deep fruit detection in orchards. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3626–3633. IEEE, 2017.

[11] Juan Pablo Vasconez, Jose Delpiano, Stavros Vougioukas, and F Auat Cheein. Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation. *Computers and Electronics in Agriculture*, 173:105348, 2020.

[12] Hamzeh Mirhaji, Mohsen Soleymani, Abbas Asakereh, and Saman Abdanan Mehdizadeh. Fruit detection and load estimation of an orange orchard using the yolo models through simple approaches in different imaging and illumination conditions. *Computers and Electronics in Agriculture*, 191:106533, 2021.

[13] Craig B MacEachern, Travis J Esau, Arnold W Schumann, Patrick J Hennessy, and Qamar U Zaman. Detection of fruit maturity stage and yield estimation in wild blueberry using deep learning convolutional neural networks. *Smart Agricultural Technology*, 3:100099, 2023.

[14] Lei Shen, Jinya Su, Runtian He, Lijie Song, Rong Huang, Yulin Fang, Yuyang Song, and Baofeng Su. Real-time tracking and counting of grape clusters in the field based on channel pruning with yolov5s. *Computers and Electronics in Agriculture*, 206:107662, 2023.

[15] Dandan Wang and Dongjian He. Channel pruned yolo v5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosystems Engineering*, 210:271–281, 2021.

[16] G. Jocher, A. Chaurasia, and J. Qiu. Yolo by ultralytics. https://github.com/ultralytics/ultralytics, 2023.

[17] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

[18] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

[19] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors. *arXiv preprint arXiv:2411.00201*, 2024.

[20] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[21] Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

[22] Baoling Ma, Zhixin Hua, Yuchen Wen, Hongxing Deng, Yongjie Zhao, Liuru Pu, and Huaibo Song. Using an improved lightweight yolov8 model for real-time detection of multi-stage apple fruit in complex orchard environments. *Artificial Intelligence in Agriculture*, 11:70–82, 2024.

[23] Aobin Zhu, Ruirui Zhang, Linhuan Zhang, Tongchuan Yi, Liwan Wang, Danzhu Zhang, and Liping Chen. Yolov5s-cedb: A robust and efficiency camellia oleifera fruit detection algorithm in complex natural scenes. *Computers and Electronics in Agriculture*, 221:108984, 2024.

[24] Wenbai Chen, Mengchen Liu, ChunJiang Zhao, Xingxu Li, and Yiqun Wang. Mtd-yolo: Multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. *Computers and Electronics in Agriculture*, 216:108533, 2024.

[25] Jiuxin Wang, Man Liu, Yurong Du, Minghu Zhao, Hanlang Jia, Zhou Guo, Yaoheng Su, Dingze Lu, and Yucheng Liu. Pg-yolo: An efficient detection algorithm for pomegranate before fruit thinning. *Engineering Applications of Artificial Intelligence*, 134:108700, 2024.

[26] Fan Meng, Jinhui Li, Yunqi Zhang, Shaojun Qi, and Yunchao Tang. Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks. *Computers and Electronics in Agriculture*, 214:108298, 2023.

[27] Xueyan Zhu, Fengjun Chen, Xinwei Zhang, Yili Zheng, Xiaodan Peng, and Chuang Chen. Detection the maturity of multi-cultivar olive fruit in orchard environments based on olive-efficientdet. *Scientia Horticulturae*, 324:112607, 2024.

[28] Tian Zhang, Dongfang Zhao, Yesheng Chen, Hongli Zhang, and Shulin Liu. Deepsort with siamese convolution autoencoder embedded for honey peach young fruit multiple object tracking. *Computers and Electronics in Agriculture*, 217:108583, 2024.

[29] Thiago T Santos, Kleber XS de Souza, João Camargo Neto, Luciano V Koenigkan, Alécio S Moreira, and Sônia Ternes. Multiple orange detection and tracking with 3-d fruit relocalization and neural-net based yield regression in commercial sweet orange orchards. *Computers and Electronics in Agriculture*, 224:109199, 2024.

[30] Arturo Aquino, Juan Manuel Ponce, and José Manuel Andújar. Identification of olive fruit, in intensive olive orchards, by means of its morphological structure using convolutional neural networks. *Computers and Electronics in Agriculture*, 176:105616, 2020.

[31] Hatice Catal Reis and Veysel Turk. Potato leaf disease detection with a novel deep learning model based on depthwise separable convolution and transformer networks. *Engineering Applications of Artificial Intelligence*, 133:108307, 2024.

[32] João Mendes, José Lima, Lino Costa, Nuno Rodrigues, and Ana I Pereira. Deep learning networks for olive cultivar identification: A comprehensive analysis of convolutional neural networks. *Smart Agricultural Technology*, 8:100470, 2024.

[33] Hai-Binh Le, Thai Dinh Kim, Manh-Hung Ha, Anh Long Quang Tran, Duy-Thuc Nguyen, and Xuan-Minh Dinh. Robust surgical tool detection in laparoscopic surgery using yolov8 model. In *2023 International Conference on System Science and Engineering (ICSSE)*, pages 537–542. IEEE, 2023.

[34] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[35] Manh-Tuan Do, Manh-Hung Ha, Duc-Chinh Nguyen, Kim Thai, and Quang-Huy Do Ba. Human detection based yolo backbones-transformer in uavs. In *2023 International Conference on System Science and Engineering (ICSSE)*, pages 576–580. IEEE, 2023.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[37] Jinsheng Xiao, Tao Zhao, Yuntao Yao, Qiuze Yu, and Yunhua Chen. Context augmentation and feature refinement network for tiny object detection, 2022. Retrieved from https://openreview.net/forum?id=q2ZaVU6bEsT.

[38] Yi-Fan Zhang, Weiqiang Ren, Zhang Zhang, Zhen Jia, Liang Wang, and Tieniu Tan. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing*, 506:146–157, 2022.

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding

box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[40] Zanjia Tong, Yuhang Chen, Zewei Xu, and Rong Yu. Wise-iou: Bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*, 2023.

[41] Zhora Gevorgyan. Siou loss: More powerful learning for bounding box regression. *arXiv preprint arXiv: 2205.12740,* 2022.