

Article

Vision-Based Strawberry Ripeness Grading for Harvest Decision Support in Smart Greenhouses

Gwanghyeong Lee , Deepak Ghimire , Sewoon Cho , Jihwan Seo, Sunghwan Jeong and Byoungjun Kim * 

AI Application Research Center, Korea Electronics Technology Institute, Jeonju 54853, Republic of Korea; lightbro@keti.re.kr (G.L.); deepak@keti.re.kr (D.G.); swcho@keti.re.kr (S.C.); seojh410@keti.re.kr (J.S.); shjeong@keti.re.kr (S.J.)

* Correspondence: jun0420@keti.re.kr

Abstract

Strawberry ripeness estimation is essential for determining optimal harvest timing and supporting quality management in smart greenhouse environments. However, since strawberry ripening is a continuous process rather than a transition between discrete visual boundaries, standard RGB-based methods often struggle to distinguish subtle color transitions during intermediate ripeness levels, and misclassifications among adjacent ripeness stages can substantially affect harvest decisions. This makes it important to not only achieve high classification accuracy but also to minimize the ordinal deviation between predicted and actual ripeness stages. To address these limitations, this study presents a two-stage vision-based framework for five-class strawberry ripeness grading. In the first stage, a YOLO11m-based detector localizes strawberry regions within greenhouse images and generates regions of interest (ROIs). In the second stage, each extracted ROI is classified using a five-channel input representation that combines RGB, Excess Green (ExG), and Hue information to enhance color-sensitive feature extraction. In addition, a specialized backbone, termed CP2-A2C2F, is designed to preserve high-resolution P2 features and integrate them with deeper P4 semantic representations through a dual-scale fusion strategy. This structure enables the model to capture localized chromatic variations and contextual maturity information simultaneously. Experimental results show that the CP2-A2C2F model achieved the best overall classification accuracy of 87.7% and a weighted F1 score of 0.877, together with the lowest mean absolute error (MAE) of 0.162 among models equipped with an auxiliary regression head, indicating the smallest average ordinal deviation between predicted and ground-truth ripeness stages. The ablation study further confirmed that adding ExG and Hue channels moderately reduced ordinal prediction error compared with the RGB-only configuration, and that the proposed P2 + P4 fusion achieved the best accuracy, weighted F1 score, and MAE among the evaluated fusion configurations, indicating that direct fusion of high-resolution P2 features with deeper P4 semantic features provides a more effective balance than including the intermediate P3 level. These results indicate that the combination of explicit color representation and dual-scale P2 + P4 feature fusion can improve five-class strawberry ripeness grading and help reduce ordinal prediction errors under the tested greenhouse conditions.

Keywords: strawberry ripeness classification; smart greenhouse; YOLO11m; RGB image; Excess Green; Hue; CP2-A2C2F; computer vision; harvest decision support



Academic Editor: Luca Ghiani

Received: 10 June 2026

Revised: 17 July 2026

Accepted: 17 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Strawberries are widely cultivated and consumed worldwide in various cultivars and are recognized not only for their characteristic flavor and aroma but also for their antioxidant compounds and associated health benefits [1–3]. Strawberry ripeness is a critical factor that strongly influences fruit quality, flavor, and consumer acceptance. In particular, strawberry ripeness is closely related to harvest timing, quality maintenance, distribution stability, and marketability [4–6]. In smart farms and commercial greenhouse environments, the demand for automation technologies capable of producing agricultural products with consistent quality has continuously increased. Strawberry maturity is generally determined based on color changes on the fruit surface, which is directly associated with marketability evaluation. However, in practical cultivation environments, stable ripeness estimation remains challenging due to various environmental disturbances such as illumination variation, partial occlusion by foliage, complex background structures, and overlapping fruits. Beyond these environmental factors, a more fundamental challenge lies in the continuous nature of the ripening process itself: since ripeness progresses gradually rather than through discrete visual boundaries, an effective harvest decision support system must not only assign a ripeness class but also minimize the ordinal distance between predicted and actual maturity stages when misclassification occurs.

Traditional strawberry ripeness discrimination has mainly relied on manual visual inspection performed by workers. Although this approach is intuitive, the inspection results may vary significantly depending on the experience and subjectivity of the evaluator. Furthermore, in large scale cultivation environments, manual inspection causes high labor costs and substantial time consumption. Before the widespread adoption of deep learning techniques, several studies attempted to estimate fruit ripeness using conventional image processing approaches based on handcrafted color, texture, and shape features [7–9]. Although these methods showed partial effectiveness under controlled environments, their performance was often sensitive to illumination variation and complex background conditions. To overcome these limitations, research on automated maturity classification based on computer vision and deep learning has been actively conducted in recent years. In particular, convolution-based classification models and object detection-based frameworks have demonstrated strong performance in agricultural image analysis and have been applied to various crop monitoring tasks [10].

Recent studies have proposed methods for automatically classifying fruit maturity by combining YOLO-based object detection models with CNN or vision transformer-based backbone networks [11–14]. These sequential frameworks have shown strong object detection performance under practical greenhouse conditions. However, most existing methods rely on standard RGB images, which makes it difficult to distinguish subtle and localized color transitions during intermediate ripening stages, such as early turning, turning, and late turning.

To address the limitations of RGB-based representations, several studies have explored alternative imaging modalities and spectral information for strawberry ripeness assessment [15–17]. Near infrared (NIR), mid infrared (MIR), and vegetation index-based features can provide additional information on ripeness-related physiological and surface changes across different wavelength ranges. However, these approaches require specialized imaging sensors, which can limit their scalability and practical deployment in commercial greenhouse environments.

Therefore, improving the use of accessible color information within standard vision networks remains a practical direction for strawberry ripeness classification. In conventional CNN backbones, repeated downsampling can reduce high-resolution color and texture cues that are important for identifying localized pigmentation on the fruit surface.

A feature representation that preserves fine scale chromatic information while maintaining global semantic context is therefore required to better capture the color-based cues needed for fine-grained ripeness discrimination, without relying on additional imaging hardware.

In addition to these general approaches, three related research directions provide further context for the present study. First, several works have combined RGB imagery with vegetation index-based representations to enhance fine grained visual discrimination in agricultural classification tasks for instance, new RGB-based vegetation indices have been combined with CNN-based models to identify viral symptoms on pepper leaves, illustrating the potential benefit of explicit color derived features over RGB alone [18]. This line of work has mainly been explored in the context of disease identification, while its application to continuous ripeness grading remains relatively limited. Second, multi-scale feature fusion strategies have been actively investigated for fine grained agricultural classification, including approaches that combine multi-scale feature fusion with attention mechanisms for crop pest classification [19] and methods that integrate improved ResNet backbones with BiFPN-based multi-scale fusion for fruit recognition [20]. These studies highlight the value of multi-scale fusion for capturing subtle visual differences, and similar principles may also be relevant for preserving high-resolution chromatic information in continuous ripeness transitions. Third, ordinal regression has been explored as a way to model the inherent ordering among quality or ripeness grades, such as an ordinal regression-based extreme learning machine method for apple grading [21]. Recent degraded-vision restoration work, such as the lightweight multi-degradation UFR-GAN model, also highlights a complementary direction for improving visual inputs before downstream detection or recognition tasks under degraded imaging conditions [22]. Building on these complementary directions, the present study aims to combine explicit color-based feature representation, high-resolution preserving multi-scale fusion, and ordinal-aware learning within a single framework for fine-grained strawberry ripeness grading.

To address these challenges, this study proposes a two-stage strawberry ripeness grading framework. In the first stage, strawberry regions are localized within complex greenhouse environments using a YOLO11m-based object detection model [23]. In the second stage, five-class ripeness grading is performed using the extracted regions of interest (ROIs). The overall architecture of the proposed framework is illustrated in Figure 1. To improve color discrimination capability without requiring expensive hardware, a five-channel input representation based on RGB + ExG + Hue was constructed by extending the conventional RGB input with Excess Green (ExG) and Hue channels. The ExG channel emphasizes green dominant characteristics appearing in the initial ripeness classes, while the Hue channel provides structured color transition information throughout the continuous ripening process.

Additionally, this study introduces the CP2-A2C2F (ConcatP2-A2C2F) backbone architecture, which is inspired by the A2C2F module used in recent YOLOv12-based architectures [24] and redesigned for fine-grained strawberry ripeness classification. Unlike general object detection backbones, the CP2-A2C2F architecture explicitly preserves the high-resolution P2 feature map and integrates it with deeper P4 semantic representations through a dual scale fusion strategy. This design is intended to reduce the loss of localized chromatic information caused by repeated downsampling, thereby maintaining high-resolution texture features and local color representations. By combining shallow color-sensitive features with deeper contextual information, the proposed framework can better capture subtle color differences appearing in intermediate ripeness classes under the tested greenhouse conditions.

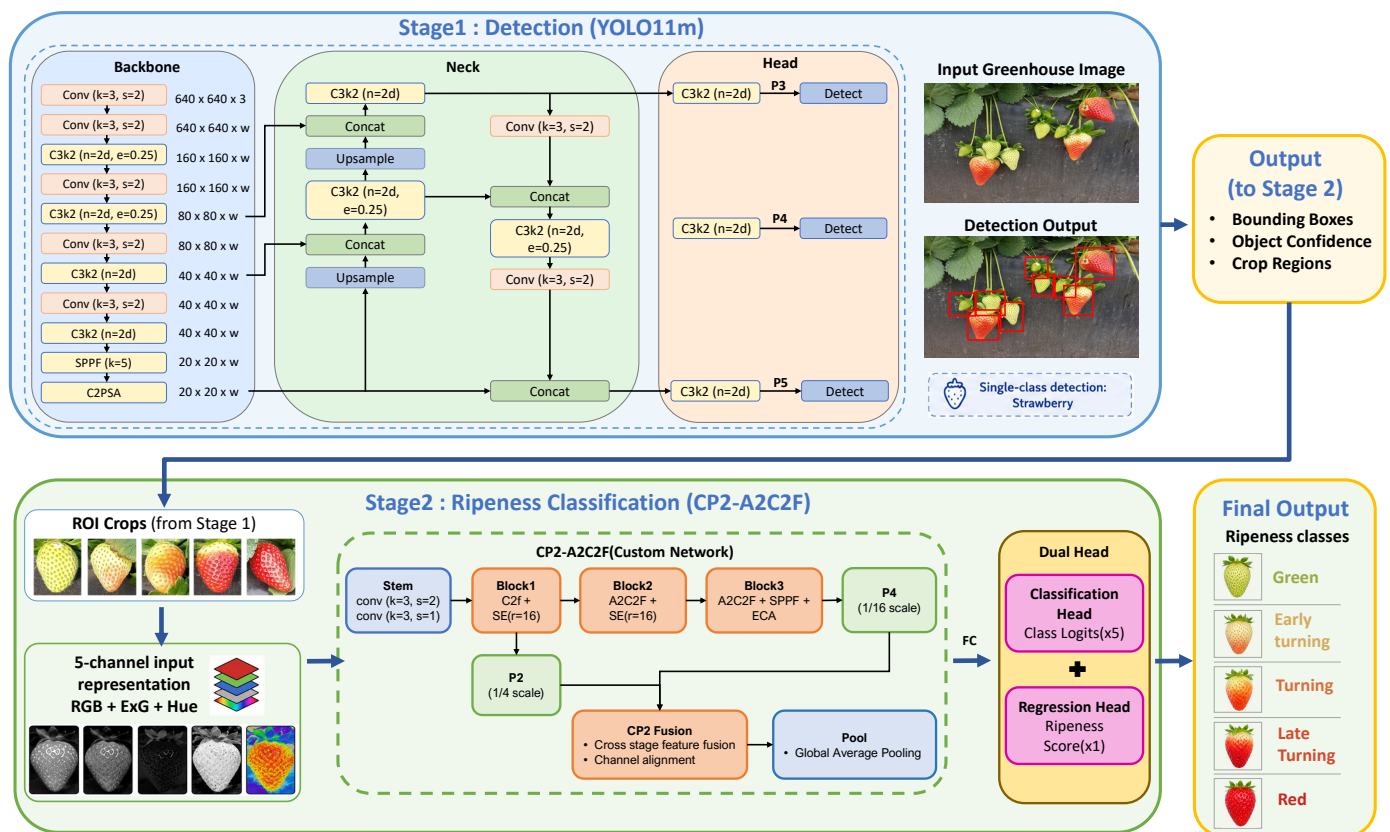


Figure 1. Overall architecture of the proposed two-stage strawberry ripeness grading framework.

2. Materials and Methods

2.1. Building Strawberry Maturity Dataset

Deep learning-based approaches have been widely applied to fruit ripeness classification and have shown strong performance in distinguishing maturity classes. However, the effectiveness of these methods depends not only on model design but also on the availability of sufficiently large and well annotated datasets.

Publicly available large scale vision datasets such as ImageNet, COCO, and Objects365 have contributed significantly to general image classification and object detection research [25–27]. However, these datasets are not specifically designed for fine-grained strawberry ripeness classification. In particular, images that represent gradual color transitions between adjacent maturity classes are scarce, making it difficult to train models that can distinguish subtle differences among intermediate ripeness classes.

Therefore, in this study, a dedicated strawberry maturity dataset was constructed to support reliable model training and evaluation across five ripeness classes.

2.1.1. Image Collection

The dataset was collected from a single commercial greenhouse facility located in Pyeongtaek, Gyeonggi Province, Republic of Korea, during the 2025 growing season. Images were captured using a Samsung Galaxy S22 Ultra smartphone (Samsung Electronics Co., Ltd., Suwon, Republic of Korea) equipped with a high-resolution CMOS image sensor suitable for field image acquisition. The study focused on the Seolhyang strawberry cultivar, which is the dominant cultivar in South Korea. According to the Korea Rural Economic Institute, Seolhyang accounted for 80.6% of the domestic strawberry cultivation area in the 2024/2025 season [28]. This cultivar exhibits distinct color changes during the maturation process, making it suitable for visual ripeness classification.

To represent the tested greenhouse acquisition conditions, data were acquired under diverse lighting environments, ranging from direct sunlight to overcast conditions. The dataset reflects common greenhouse challenges, including variations in illumination, partial occlusions caused by foliage, and complex background interference from stems and overlapping leaves. During the data quality control process, images with severe motion blur caused by camera shake were manually excluded. A total of approximately 3000 images were collected, covering the full range of strawberry maturity classes. Representative samples of the collected dataset are presented in Figure 2.

It should be noted that the dataset was collected from a single commercial greenhouse facility using one specific strawberry cultivar (Seolhyang) and one mobile imaging device. While this setup reflects a realistic deployment scenario, the limited diversity in acquisition conditions may constrain the generalizability of the trained models to other greenhouse environments, cultivars, or imaging platforms.



Figure 2. Representative samples of the collected strawberry dataset.

2.1.2. Ripeness Class Definition

In previous strawberry developmental studies, fruit maturity has often been divided into multiple developmental categories, such as green, white, turning, red, ripe, and over-ripe categories, based on visible changes in fruit surface color and texture [29,30]. However, the objective of this study is not physiological developmental analysis, but image-based ripeness classification under the tested greenhouse conditions. Therefore, the strawberry samples were categorized into five visually distinguishable ripeness classes based on the proportion and distribution of red coloration on the fruit surface: green, early turning, turning, late turning, and red.

The class criteria were defined according to the visible proportion of red coloration on the fruit surface, following the general use of surface color as a maturity indicator in strawberry quality assessment. Representative examples and characteristics for each class are summarized in Table 1. The green class corresponds to unripe strawberries with greenish or whitish surfaces and no visible red pigmentation. The early turning class represents the initial appearance of red areas, typically covering a small portion of the fruit surface. The turning class indicates a transitional phase in which red and green regions coexist, reflecting an intermediate ripening process.

The late turning class is characterized by a predominantly red surface, although small areas of lighter or pale green coloration may remain. The red class corresponds to fully ripe strawberries with uniform red coloration across almost the entire fruit surface. This five-class definition provides a practical and consistent criterion for distinguishing strawberry ripeness levels in this study.

To verify the reliability of the annotation process, an inter annotator agreement study was conducted on a subset of 1000 ROI samples. Two additional annotators independently labeled the samples without access to the original ground truth labels, resulting in three independent label sets including the original ground truth. The overall Fleiss' kappa across all three label sets was 0.810, indicating almost perfect agreement, with an average pairwise agreement rate of 86.17% across the three annotator pairs. Samples without full agreement

across the three sets were treated as inconsistent and re labeled through discussion among all annotators until consensus was reached.

Table 1. Definition of strawberry ripeness classes based on surface color characteristics.

Ripeness Class	Image	Feature
Green		Predominantly green or whitish-green fruit surface with no visible red pigmentation, representing the immature and non-harvestable stage.
Early turning		Small red patches begin to appear on the surface, covering approximately 10–30% of the fruit, while most areas remain green or pale.
Turning		Red coloration spreads over approximately 30–60% of the fruit surface, although green or pale regions are still clearly visible.
Late turning		Most of the fruit surface is red, covering approximately 60–90%, with only limited green or pale regions remaining.
Red		Nearly the entire fruit surface shows uniform red coloration, indicating the fully ripe and harvest-ready stage.

2.2. Two-Stage Framework Overview

The proposed framework consists of a sequential two-stage pipeline designed to perform strawberry localization and ripeness grading in greenhouse environments. The overall architecture is illustrated in Figure 1.

In the first stage, a YOLO11m-based detector is employed to identify strawberry regions from high-resolution greenhouse images. This stage provides strawberry localization for region of interest (ROI) generation under the tested greenhouse conditions. After detection, each strawberry region is cropped as a ROI and transferred to the second stage for ripeness grading.

In the second stage, the CP2-A2C2F model predicts the ripeness class of each extracted strawberry ROI. To better represent chromatic changes associated with strawberry maturation, the conventional RGB image is expanded into a five-channel representation by incorporating Excess Green (ExG) and Hue information. The ExG channel emphasizes green dominant characteristics that frequently appear in early ripeness classes, while the Hue channel provides additional color transition information related to gradual maturation patterns.

The CP2-A2C2F architecture preserves high-resolution P2 features while simultaneously utilizing deeper semantic representations extracted from the P4 feature level. These multi-scale features are integrated through the CP2 Fusion module, which combines shallow chromatic information with high-level contextual representations. Finally, the aggregated features are processed through dual prediction heads for ripeness classification and regression-based ripeness estimation. By combining explicit color representations with multi-scale feature fusion, the framework aims to improve fine-grained ripeness grading in this setting. Table 2 summarizes the correspondence between each architectural component and the specific ripeness grading challenge it is designed to address.

Table 2. Correspondence between architectural components and the ripeness grading challenges they address.

Component	Targeted Challenge	Related Evidence
ExG + Hue	RGB alone may insufficiently represent gradual color transitions during early and intermediate ripening stages.	Section 3.4
P2–P4 fusion	Downsampling can reduce localized chromatic and texture cues needed to distinguish adjacent intermediate classes.	
Regression head	Independent class treatment does not explicitly reflect the ordinal relationship among ripeness stages.	

2.3. Five-Channel Input Representation

To enhance ripeness discrimination, a five-channel input representation is constructed by extending the conventional RGB input with additional color-based features.

Specifically, two additional channels, Excess Green (ExG) and Hue, are incorporated to better capture color variations associated with strawberry maturation. Strawberry ripening is characterized by a gradual transition from green to red coloration on the fruit surface.

The ExG channel is calculated using the following linear combination of the RGB color components [31]:

$$\text{ExG} = 2G - R - B \quad (1)$$

where R , G , and B denote the red, green, and blue color components, respectively, each normalized to the range $[0, 1]$ prior to computation.

This representation emphasizes the green channel and was designed to highlight regions where green coloration is dominant, which is characteristic of early ripeness stages. ExG thus provides an explicit color-derived cue intended to represent the transition from green-dominant fruit surfaces toward later ripeness stages.

The Hue channel is derived from the HSV color space and represents the dominant chromatic component of each pixel. In this study, the Hue channel was obtained using the OpenCV RGB-to-HSV conversion, where the hue component is represented in the range of 0 to 179. Hue provides a structured representation of chromatic transitions from green to red and is intended to complement ExG when modeling gradual color changes across intermediate ripeness classes such as turning and late turning.

Before constructing the five-channel input, the RGB image was converted to floating point values and scaled to the range of $[0, 1]$. The RGB channels were then normalized using the ImageNet mean and standard deviation values. The ExG values were then computed from the scaled RGB channels and normalized using a mean of 0.0 and a standard deviation of 0.5. The Hue values obtained from the OpenCV HSV representation were scaled by 255 and then normalized using a mean of 0.35 and a standard deviation of 0.20.

Finally, the normalized RGB, ExG, and Hue channels were concatenated in the order of R , G , B , ExG, and Hue to form a five-channel tensor. By combining these complementary color representations, the model is designed to better capture subtle differences across ripeness classes.

2.4. CP2-A2C2F Backbone Architecture

To improve feature representation for fine-grained strawberry ripeness classification, a modified backbone architecture, termed CP2-A2C2F, was developed. The CP2-A2C2F backbone is inspired by the A2C2F module used in YOLOv12 [24], but is redesigned for ripeness classification rather than general object detection. In particular, this archi-

texture focuses on preserving localized color and texture information extracted from the RGB + ExG + Hue five-channel representation, as illustrated in Figure 3.

Unlike general object recognition tasks that mainly rely on high level semantic features, strawberry ripeness classification depends strongly on subtle local color transitions distributed across the fruit surface. This is especially important for intermediate ripeness classes, such as early turning, turning, and late turning, where only partial red coloration appears. If these local color patterns are weakened by repeated downsampling, the model may lose information that is important for distinguishing adjacent ripeness classes.

This limitation becomes more important when using the RGB + ExG + Hue five-channel representation, as the additional ExG and Hue channels contain fine scale chromatic variations associated with strawberry maturation. Therefore, the backbone is designed to preserve high-resolution feature maps sufficiently to utilize these additional color channels by directly connecting high-resolution P2 features to the final fusion module.

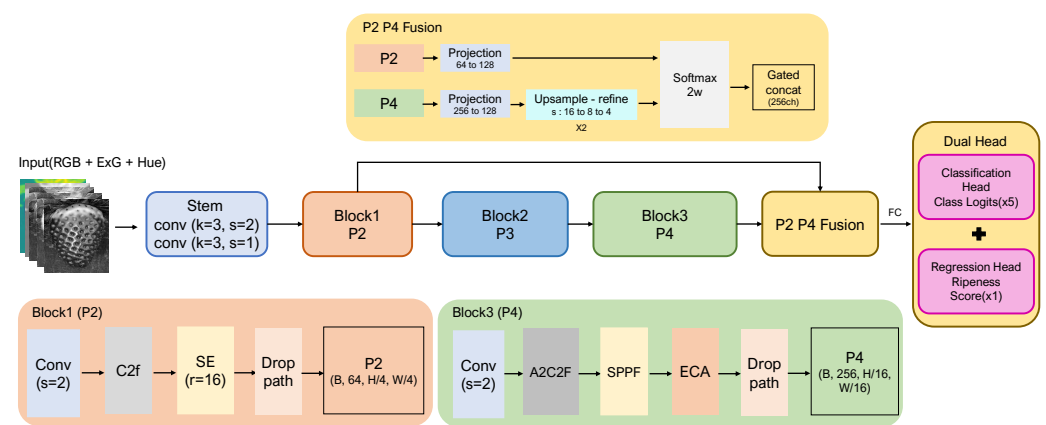


Figure 3. Overall structure of the CP2-A2C2F backbone architecture.

2.4.1. Hierarchical Feature Extraction

The CP2-A2C2F backbone consists of a stem module, three hierarchical feature extraction blocks, a dual scale feature fusion module, and dual prediction heads. The stem module is composed of two consecutive ConvBNAct blocks, each consisting of a convolution layer, batch normalization, and SiLU activation. This module performs initial low level feature extraction from the five-channel input representation.

After the stem module, the first feature extraction block generates the P2 feature map. This block consists of a downsampling layer, a C2f block, and an SEBlock, together with a light DropPath rate of 0.03. Since P2 maintains a relatively high spatial resolution, it contains fine local color and texture information that is important for detecting subtle ripeness cues on the strawberry surface. The C2f block enables efficient feature reuse through bottleneck-based feature aggregation, while the SEBlock recalibrates channel responses to emphasize informative chromatic features [32].

The second feature extraction block generates the P3 feature map using an A2C2F-based feature extraction block. The P3 feature map represents an intermediate level feature representation between the high-resolution P2 feature and the deeper P4 feature. Although P3 is not directly connected to the final CP2 fusion module, it serves as an intermediate transformation block that progressively abstracts ripeness-related features before the third feature extraction block.

The third feature extraction block produces the P4 feature map, which contains deeper semantic information related to the overall maturity appearance of the strawberry. This block employs an A2C2F block, followed by an SPPF module and an ECA block, together with a DropPath rate of 0.08. The SPPF module expands the receptive field to capture

broader contextual information, while the ECA block performs lightweight channel attention to strengthen informative channels without introducing a large number of additional parameters [33].

The internal structures of the C2f and A2C2F blocks used in this backbone are shown in Figure 4. The C2f block first splits the input features and then processes part of the feature channels through repeated bottleneck operations. The processed features are subsequently concatenated and refined using a 1×1 convolution, enabling efficient feature reuse while maintaining a lightweight aggregation process. In contrast, the A2C2F block replaces the standard bottleneck units with attention enhanced bottleneck modules. This design allows the network to refine ripeness-related feature responses more effectively while preserving the efficient channel splitting and concatenation mechanism.

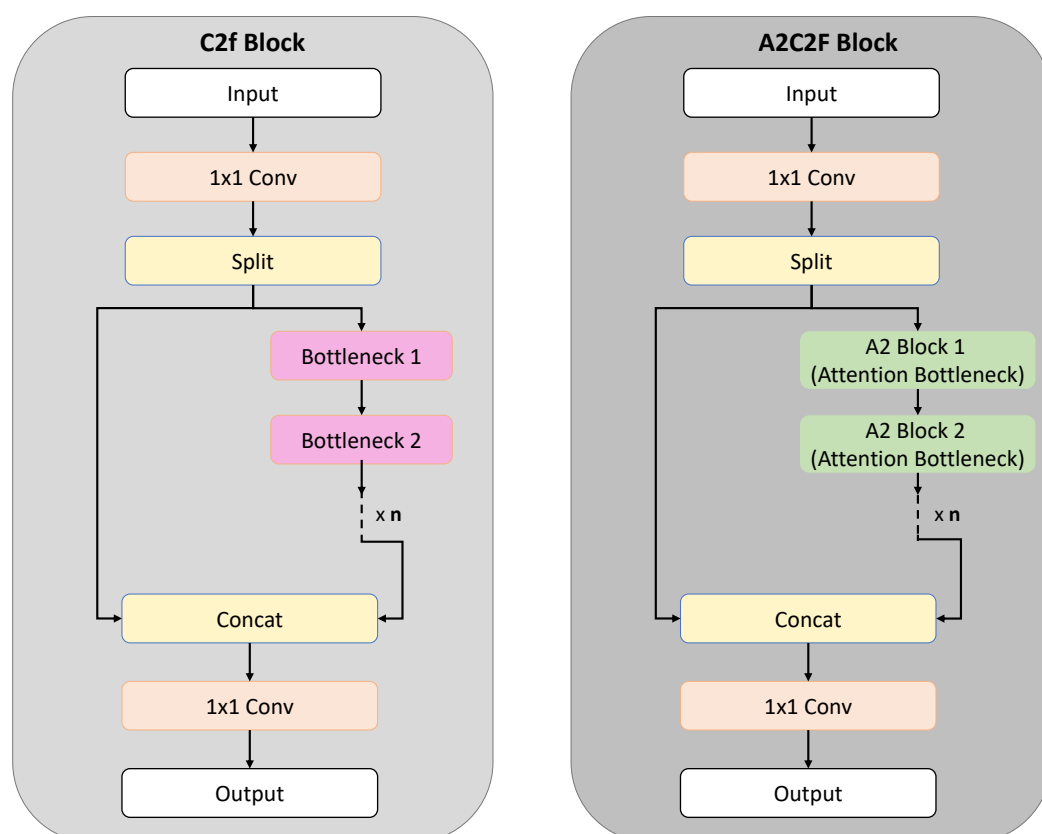


Figure 4. Internal structures of the C2f and A2C2F blocks used in the CP2-A2C2F backbone.

Through this hierarchical feature extraction process, the CP2-A2C2F backbone obtains complementary feature representations: P2 preserves high-resolution local chromatic and texture cues, P3 provides intermediate abstraction, and P4 captures high level semantic maturity information. Among these, P2 and P4 are directly utilized in the final CP2 fusion module to combine localized color details with global ripeness context.

Table 3 summarizes the scale and main role of each hierarchical feature map in the CP2-A2C2F backbone.

Table 3. Summary of hierarchical feature maps in the CP2-A2C2F backbone.

Feature Block	Feature Map	Scale	Main Role
Block 1	P2	1/4	High-resolution local color and texture preservation
Block 2	P3	1/8	Intermediate feature abstraction
Block 3	P4	1/16	High level semantic ripeness representation

2.4.2. Dual Scale Feature Fusion and Prediction Heads

The P2 feature map contains high-resolution spatial and chromatic information that captures fine-grained color and texture cues, while the P4 feature map contains deeper semantic information related to the overall appearance of strawberry maturity, as summarized in Table 3. Conventional multi-scale fusion architectures such as FPN [34], PANet [35], and BiFPN [36], including the fusion modules embedded in recent YOLO variants, integrate feature maps through sequential connections between adjacent pyramid levels and typically produce multiple scale specific outputs to support object detection at varying object sizes. In contrast, the CP2-A2C2F backbone targets a single-decision classification task, where only one unified ripeness representation per ROI is required rather than multiple scale specific outputs. Accordingly, instead of aggregating all hierarchical feature maps through sequential adjacent level connections, the CP2-A2C2F backbone directly fuses only the P2 and P4 levels, while omitting the intermediate P3 level from the final fusion. This decision is based on the hypothesis that the P3 level provides an intermediate abstraction that may contain less fine scale chromatic detail than P2 while providing less mature semantic context than P4. The effectiveness of this design choice is evaluated empirically in the feature fusion ablation study presented in Section 3.4.

In this implementation, the P2 and P4 feature maps are first projected to a common channel dimension of 128 using 1×1 convolutions. To bridge the four-fold resolution gap between P4 and P2, the projected P4 feature is upsampled in two sequential $2\times$ steps (stride 16 to 8, then stride 8 to 4) rather than a single $4\times$ upsampling operation. Each step is followed by a refine convolution with a residual connection and a zero-initialized batch normalization layer, together with Dropout2d(0.1) for regularization, which helps stabilize training when adding capacity to the upsample path. It should be noted that this intermediate resolution step operates only on the projected P4 feature and does not incorporate the P3 feature map produced by Block2; the actual P3-level representation therefore remains excluded from the final fused output, consistent with the design rationale described above. The aligned P2 and P4 features are then combined through a scale gate that computes pure softmax weights ($2w$) rather than a $(1 + w)$ formulation, which allows either branch to be fully suppressed when uninformative. The gated P2 and P4 features are subsequently concatenated to form a 256-channel fused representation. This design reduces redundant feature propagation while maintaining a lightweight feature aggregation structure.

The fused representation combines complementary information from two different feature levels. Specifically, P2 provides localized color and texture cues that are important for detecting early ripening patterns, whereas P4 provides global maturity context. Finally, the fused feature representation is compressed through global average pooling and transformed into a 512 dimensional shared feature vector using a fully connected layer followed by SiLU activation and Dropout(0.25). This shared feature vector is then processed by two prediction heads. The classification head consists of a fully connected layer, SiLU activation, dropout, and a final linear layer that outputs five class logits. The regression head follows the same structure but outputs a single continuous ripeness score. Table 4 summarizes the channel dimensions and cumulative downsampling strides of each module within the CP2-A2C2F backbone, including the stem, the three hierarchical feature extraction blocks, the P2 and P4 projection layers, the scale gate, and the final fusion module.

Table 4. Detailed channel dimensions and output feature map strides of the CP2-A2C2F backbone.

Module	Input Ch.	Output Ch.	Input Stride	Output Stride
Stem	5	32	1	2
Block1 (P2)	32	64	2	4
Block2 (P3)	64	128	4	8
Block3 (P4)	128	256	8	16
P2 projection	64	128	4	4
P4 projection	256	128	16	16
P4 upsample step 1 (refine conv)	128	128	16	8
P4 upsample step 2 (refine conv)	128	128	8	4
Scale gate (softmax)	128	128	4	4
CP2 fusion	256	256	4	4
Pool + FC	256	512	4	–

2.5. Experimental Setup

The proposed framework was implemented using the PyTorch version 2.9.1 [37] and trained on a Linux-based system equipped with an NVIDIA GeForce RTX 3080 GPU. Unless otherwise stated, experiments were conducted using a fixed random seed of 42 to ensure reproducibility.

The constructed dataset consisted of 3000 images and was divided into training, validation, and test sets containing 2000, 500, and 500 images, respectively. The dataset was split at the original-image level before ROI cropping. Each image represented an independent capture event, and no burst-mode or consecutive-frame sequences were included across the training, validation, and test subsets, thereby minimizing the risk of leakage from near-duplicate images. Each subset included samples from all five ripeness stages. For the first-stage detector, a total of 6665, 1848, and 2175 manually annotated strawberry instances were used for training, validation, and testing, respectively. Since the second-stage ROI crops were generated directly from these annotated bounding boxes without additional filtering, the resulting ROI sample counts were identical to the annotated instance counts for each split. Table 5 summarizes the class-wise distribution of ROI samples across the five ripeness classes for each split.

This distribution reflects a natural imbalance, as shorter intermediate stages such as turning are inherently less frequent than the green and red stages in commercial greenhouse cultivation. No class balancing technique, such as oversampling or synthetic augmentation, was applied to artificially equalize the number of samples across classes, in order to preserve the natural ripeness distribution encountered in real greenhouse environments. The validation set was used for model selection and early stopping, while the test set was used only for final performance evaluation.

Table 5. Class-wise distribution of ROI samples across training, validation, and test splits.

Split	Green	Early Turning	Turning	Late Turning	Red
Train	3574	1186	280	471	1154
Validation	1019	308	71	162	288
Test	1063	442	69	99	502

For the first stage of the framework, a YOLO11m detector initialized with pretrained weights was used for strawberry localization. The detector was trained as a single class strawberry detector, and ripeness labels were assigned only for the second-stage grading task. Input images for the detector were resized to 960×960 pixels. The detected strawberry regions were cropped as regions of interest (ROIs), resized to 224×224 pixels, and then converted into the RGB + ExG + Hue five-channel representation before being passed to the CP2-A2C2F classification network.

Different preprocessing and augmentation strategies were applied to the detection and classification stages. For the detection stage, standard YOLO-based augmentations including mosaic augmentation, random scaling, and horizontal flipping were employed. Mosaic augmentation combines multiple training images into a single composite image and has been widely used in YOLO-based object detection to improve robustness under diverse object scales and backgrounds [38]. For the classification stage, color-based augmentations were intentionally excluded to preserve the original color characteristics associated with strawberry ripeness, since transformations that alter brightness, saturation, or hue could introduce signals that conflict with the ripeness labels.

The CP2-A2C2F model was trained using both ripeness classification and auxiliary regression objectives. For the auxiliary regression task, the five ripeness classes were encoded as ordinal ripeness scores. Specifically, green, early turning, turning, late turning, and red were assigned scores of 0.0, 1.0, 2.0, 3.0, and 4.0, respectively. This ordinal score representation allows the regression head to reflect the sequential progression of strawberry maturation.

The total loss function is defined as:

$$L = L_{cls} + \lambda_{reg} L_{reg} \quad (2)$$

where L_{cls} denotes the classification loss, L_{reg} denotes the auxiliary regression loss, and λ_{reg} controls the contribution of the auxiliary regression objective. In this study, L_{cls} was implemented using cross entropy loss with a label smoothing factor of 0.05, while L_{reg} was computed as the Smooth L1 loss between the predicted ripeness score $\hat{y}_j$ and the ordinal ground truth ripeness score y_j , consistent with the notation used in Equation (6). In all experiments, λ_{reg} was set to 0.25.

Key parameter selections used in the main benchmark experiments are summarized in Table 6. Among these settings, only the regression loss weighting coefficient was evaluated through a dedicated sensitivity analysis in this study; the remaining parameters were kept fixed across backbone comparisons to maintain a consistent evaluation protocol rather than to claim global hyperparameter optimality.

Table 6. Training parameter settings used for the CP2-A2C2F ripeness classification model.

Parameter	Value
Epochs	150
Batch size	32
Optimizer	AdamW [39]
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Early stopping patience	30
Random seed	42

2.6. Evaluation Metrics

Classification accuracy, weighted precision, weighted recall, weighted F1 score, and mean absolute error (MAE), where applicable, were used for performance evaluation. Weighted metrics were adopted because the number of samples may differ across ripeness classes in practical greenhouse image datasets. Unlike simple accuracy, weighted metrics reflect the contribution of each class according to its support, providing a more representative evaluation of overall classification performance.

The weighted precision, weighted recall, and weighted F1 score are defined as follows:

$$\text{Weighted Precision} = \sum_{i=1}^C \frac{n_i}{N} \text{Precision}_i \quad (3)$$

$$\text{Weighted Recall} = \sum_{i=1}^C \frac{n_i}{N} \text{Recall}_i \quad (4)$$

$$\text{Weighted F1} = \sum_{i=1}^C \frac{n_i}{N} F1_i \quad (5)$$

where C denotes the number of classes, n_i denotes the number of samples in class i , and N denotes the total number of samples. Precision_i , Recall_i , and $F1_i$ represent the precision, recall, and F1 score of class i , respectively.

For the auxiliary regression output, MAE was used to measure the absolute difference between the predicted ripeness score and the ordinal ground truth ripeness score:

$$\text{MAE} = \frac{1}{N} \sum_{j=1}^N |y_j - \hat{y}_j| \quad (6)$$

where y_j and $\hat{y}_j$ denote the ordinal ground truth ripeness score and predicted ripeness score for sample j , respectively. Since the ripeness classes were encoded as ordinal scores from 0.0 to 4.0, the MAE represents the average class-level prediction error. Therefore, a lower MAE indicates that the predicted ripeness score is closer to the true maturity class in terms of the ordinal ripening sequence.

3. Results and Discussion

The results and discussion are organized around the two stages of the proposed framework. First, the strawberry localization performance of the YOLO11m detector is evaluated to verify the reliability of ROI generation. Second, the effectiveness of the RGB + ExG + Hue five-channel input representation and the CP2-A2C2F backbone is analyzed for five-class strawberry ripeness grading.

3.1. First-Stage Strawberry Detection Performance

Before evaluating the second-stage ripeness grading model, the first-stage detector was assessed to verify whether reliable regions of interest could be generated under the tested greenhouse conditions. The YOLO11m detector was trained to localize strawberries in complex greenhouse scenes containing partial occlusion, scale variation, overlapping fruits, and background interference from leaves, stems, and cultivation structures.

Table 7 summarizes the detection performance of the YOLO11m model on both the validation and test sets. The reported performance was obtained from the best checkpoint selected during training. The detector showed comparable localization performance across the validation and test splits. In particular, the high recall values suggest that most strawberry instances were successfully detected, which is important because missed detections in the first stage cannot be recovered by the second-stage classifier.

Representative detection examples are shown in Figure 5. The detector successfully identified strawberries across various ripeness classes and greenhouse backgrounds, including small fruits and partially occluded fruits. These results suggest that the first stage can provide suitable cropped strawberry ROIs for the second-stage CP2-A2C2F ripeness classifier within this pipeline.

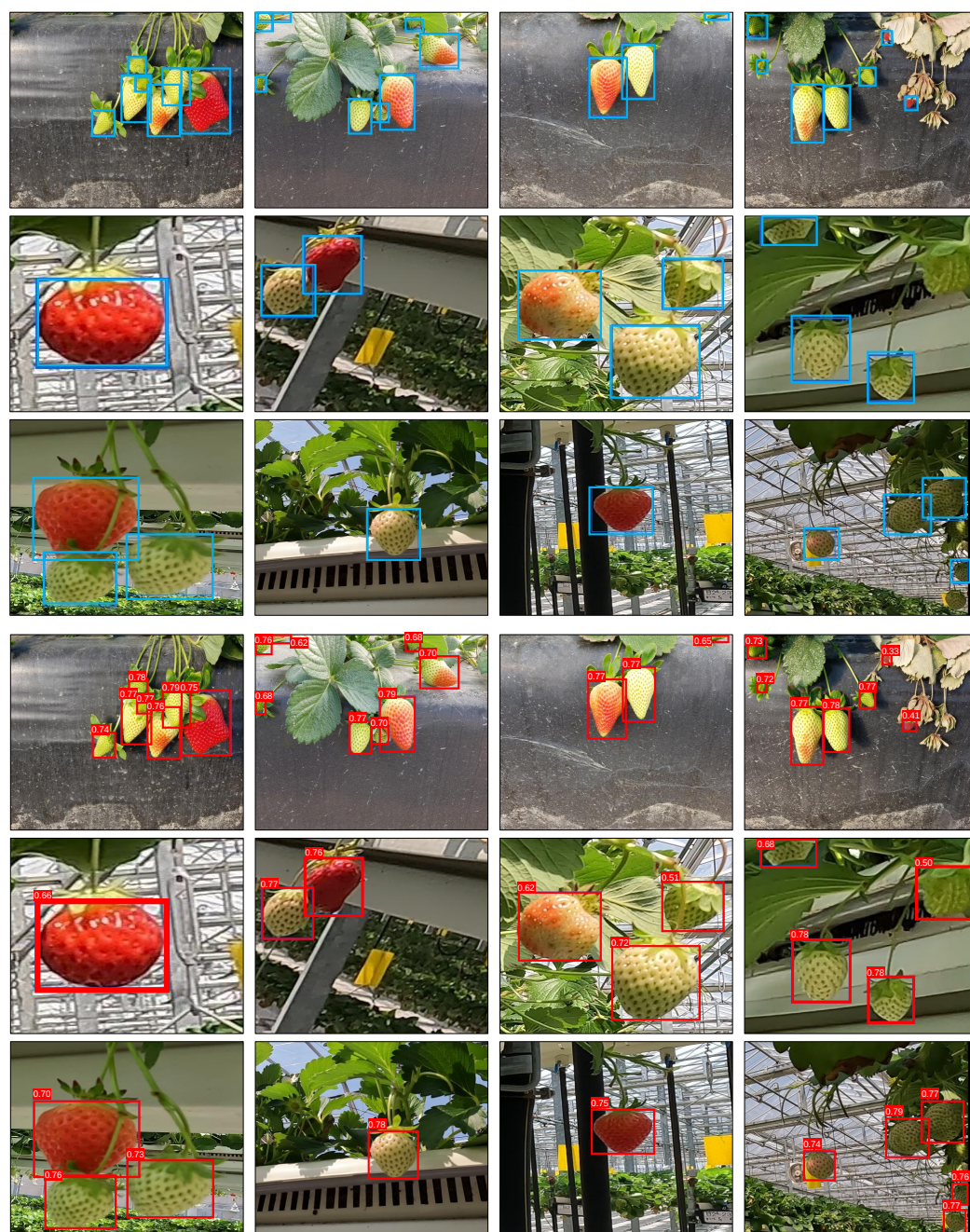


Figure 5. Representative examples for the first-stage strawberry detector. **Top:** ground-truth bounding boxes shown in blue. **Bottom:** predicted bounding boxes generated by YOLO11m shown in red.

Table 7. Detection performance of the first-stage YOLO11m strawberry detector.

Model	Split	Precision	Recall	mAP50	mAP50-95
YOLO11m [23]	Validation	0.882	0.870	0.946	0.803
	Test	0.857	0.895	0.910	0.767

3.2. Ripeness Grading Performance

The framework was evaluated for five-class strawberry ripeness grading under the tested greenhouse conditions. Performance was assessed using classification accuracy, weighted precision, weighted recall, weighted F1 score, and mean absolute error (MAE).

Table 8 summarizes the ripeness grading performance of the proposed model and comparison baselines, all evaluated using the RGB + ExG + Hue five-channel input represen-

tation. Among all evaluated models, the proposed CP2-A2C2F model achieved the lowest MAE value of 0.162, reflecting the smallest average ordinal deviation from the ground-truth ripeness stage, together with the best overall accuracy of 87.7% and a weighted F1 score of 0.877.

Table 8. Performance comparison of the proposed model and comparison baselines for five-class strawberry ripeness grading.

Model	Accuracy	MAE ↓	Weighted Prec.	Weighted Recall	Weighted F1
ResNet50 [40]	0.851	0.226	0.869	0.851	0.853
MobileNetV3 [41]	0.861	0.191	0.872	0.861	0.862
ConvNeXtV2 [42]	0.874	0.187	0.874	0.874	0.870
SwinV2 [43]	0.865	0.185	0.871	0.865	0.864
A2C2F [24]	0.865	0.198	0.874	0.865	0.867
Ours (CP2-A2C2F)	0.877	0.162	0.882	0.877	0.877

Note: Bold values indicate the best performance for each metric.

Compared with representative backbone networks such as ResNet50, MobileNetV3, ConvNeXtV2, SwinV2, and A2C2F, CP2-A2C2F demonstrated the lowest ordinal prediction error while maintaining competitive classification performance.

To further examine performance under the imbalanced class distribution of the test set, the weighted metrics reported in Table 8 were supplemented with macro-averaged and class-wise metrics. Table 9 reports macro-averaged precision, recall, and F1, balanced accuracy, and per-class precision, recall, and F1 for both the baseline A2C2F model and the proposed CP2-A2C2F model on the test set.

The per-class results show that the Turning and Late-turning categories remain more challenging than the Green and Red categories for both models. Compared with A2C2F, the proposed CP2-A2C2F model improved the F1 score for the Green class from 0.927 to 0.933, the Early-turning class from 0.764 to 0.770, the Late-turning class from 0.593 to 0.654, and the Red class from 0.912 to 0.935. In contrast, the F1 score for the Turning class decreased from 0.680 to 0.605. Macro-averaged F1 increased slightly from 0.775 to 0.779, while balanced accuracy increased from 0.799 to 0.801. These results indicate that the proposed model provided modest overall improvements, although the class-wise effect was not uniform and the Turning class remained difficult under the present test-set conditions.

Table 9. Per-class classification performance of the baseline A2C2F model and the proposed CP2-A2C2F model on the test set, together with macro-averaged metrics and balanced accuracy.

Class	Support	A2C2F (Baseline)			CP2-A2C2F (Ours)		
		Precision	Recall	F1	Precision	Recall	F1
Green	1063	0.906	0.949	0.927	0.913	0.955	0.933
Early-turning	442	0.817	0.717	0.764	0.840	0.710	0.770
Turning	69	0.630	0.739	0.680	0.527	0.710	0.605
Late-turning	99	0.500	0.727	0.593	0.602	0.717	0.654
Red	502	0.967	0.863	0.912	0.958	0.912	0.935
Macro average	2175	0.764	0.799	0.775	0.768	0.801	0.779
Balanced accuracy			0.799			0.801	

Note: Bold values indicate the better performance between the two models for each class and metric.

The results in Tables 8 and 9 were obtained from ROIs cropped using ground-truth bounding boxes so that the second-stage classifier could be evaluated independently of detection errors. An additional evaluation was therefore conducted using ROIs generated directly by the trained YOLO11m detector. Predicted boxes were generated using a confi-

dence threshold of 0.5, sorted by confidence, and matched one-to-one with ground-truth boxes using an IoU threshold of 0.5.

Of the 2175 strawberry instances in the test set, 1964 were matched with detector outputs and 211 were missed. A further 356 predicted boxes remained unmatched, corresponding to a detection precision of 0.847 and a recall of 0.903. As shown in Table 10, detector-generated ROIs produced an accuracy of 0.876, a weighted F1 score of 0.875, a macro F1 score of 0.771, and a regression-head MAE of 0.160. The corresponding values obtained from ground-truth ROIs were 0.877, 0.877, 0.779, and 0.162, respectively. The differences were small, indicating that localization errors had only a limited effect on classification for successfully detected strawberries.

Table 10. Classification performance of CP2-A2C2F using ground-truth ROIs and matched YOLO11m-generated ROIs on the test set.

Metric	Ground-Truth ROIs	Matched Detector-Generated ROIs
Accuracy	0.877	0.876
Weighted F1	0.877	0.875
Macro F1	0.779	0.771
MAE (regression head) ↓	0.162	0.160

Note: The downward arrow indicates that a lower value is better.

When missed detections were treated as failed maturity predictions, 1721 of the 2175 ground-truth instances were both detected and correctly classified, resulting in an instance-level end-to-end detection-and-correct-classification rate of 0.791. The 356 unmatched predictions also show that false detections remain a limitation of the first stage. Overall, the second-stage classifier produced similar results for ground-truth and detector-generated ROIs, whereas the complete pipeline performance was mainly affected by missed and false detections.

To examine whether the performance gain over A2C2F was consistent across different random initializations, an additional repeated-seed evaluation was conducted using three random seeds. Both A2C2F and CP2-A2C2F were evaluated under the same experimental conditions with the RGB + ExG + Hue input representation.

As shown in Table 11, CP2-A2C2F showed slightly higher average classification metrics than A2C2F, while also producing a lower MAE. The average accuracy increased from 0.867 to 0.878, corresponding to an improvement of approximately 1.1 percentage points, and the weighted F1 score increased from 0.869 to 0.878. In addition, the MAE decreased from 0.193 to 0.169 on the ordinal ripeness scale. These results indicate that CP2-A2C2F provided a moderate average improvement under the repeated-seed evaluation, with the most notable difference observed in ordinal MAE. Because the repeated-seed analysis used three random seeds, these values should be interpreted as descriptive evidence of training stability rather than as a formal statistical significance test or confidence interval analysis.

Table 11. Repeated-seed evaluation of A2C2F and the proposed CP2-A2C2F. Values are reported as mean ± standard deviation over three random seeds.

Model	Accuracy	MAE ↓	Weighted Prec.	Weighted Recall	Weighted F1
A2C2F	0.867 ± 0.008	0.193 ± 0.043	0.877 ± 0.002	0.867 ± 0.008	0.869 ± 0.008
Ours (CP2-A2C2F)	0.878 ± 0.002	0.169 ± 0.009	0.883 ± 0.001	0.878 ± 0.002	0.878 ± 0.001

Note: Bold values indicate the better mean performance between the two models for each metric.

Figure 6 presents the mean normalized confusion matrices of the baseline A2C2F and CP2-A2C2F models over three random seeds for five-class strawberry ripeness grading. As illustrated in the figure, both models achieved high recognition rates for the green and

red boundary classes, indicating that visually distinct ripeness stages were generally well distinguished based on color characteristics. Most remaining errors occurred between adjacent ripeness classes, particularly around the intermediate stages from early turning to late turning. This behavior is consistent with the continuous nature of strawberry ripening, where intermediate stages exhibit gradual color transitions rather than sharply separated visual boundaries.

The repeated-seed confusion matrix analysis showed that CP2-A2C2F improved the average recognition rate of the early turning class from 0.643 ± 0.023 to 0.723 ± 0.023 and that of the red class from 0.857 ± 0.021 to 0.910 ± 0.010 . These results suggest that preserving high-resolution P2 features and incorporating ExG and Hue information may contribute to detecting the initial appearance of red pigmentation and fully ripe red coloration more effectively. However, the class-wise improvement was not uniform: the turning class recall decreased slightly relative to A2C2F (0.78 to 0.70), and the late turning class remained comparable (0.74 to 0.73), indicating that adjacent-stage ambiguity in these intermediate classes was not fully resolved. Therefore, these results indicate that CP2-A2C2F mainly contributes to improved overall weighted performance and reduced ordinal MAE under the repeated-seed evaluation.

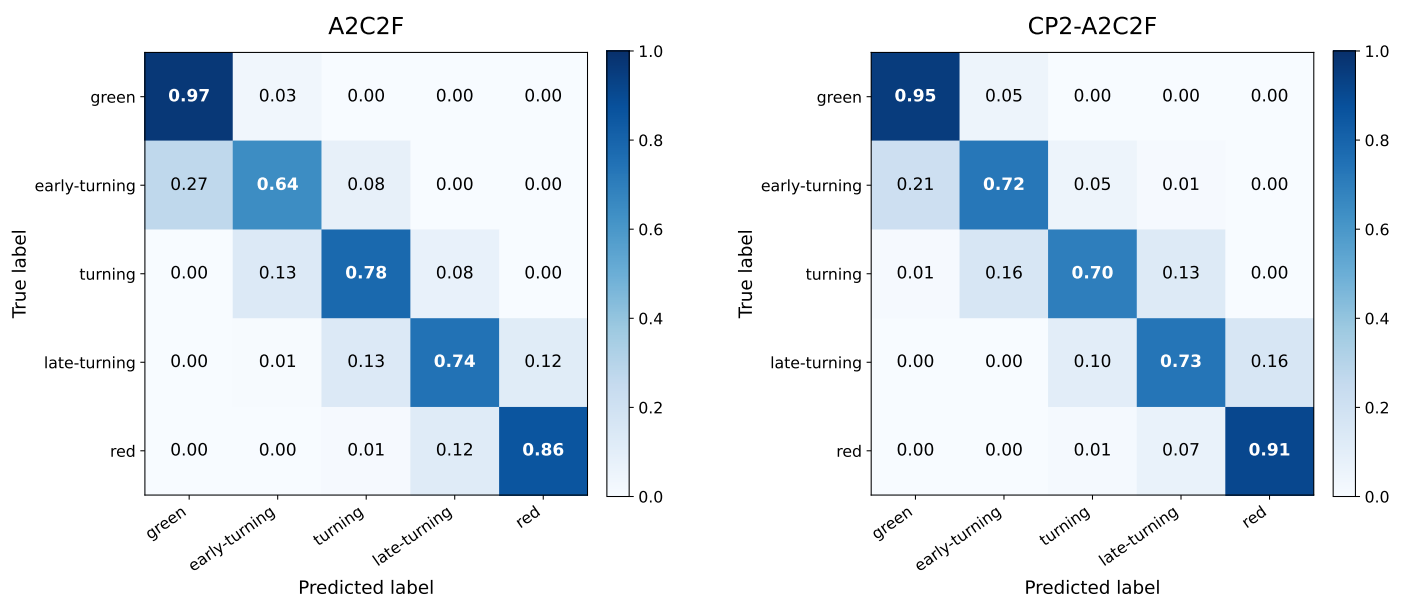


Figure 6. Mean normalized confusion matrices of A2C2F and CP2-A2C2F over three random seeds. Each cell represents the average class-normalized prediction ratio across repeated runs.

Finally, Figure 7 provides an illustrative qualitative example comparing the prediction behavior of different backbone models on the same strawberry sample. This example is intended to visualize representative prediction patterns and should be interpreted as a qualitative illustration rather than independent evidence of model superiority.

3.3. Comparison with Existing Backbones

To further interpret the results in Table 8, the influence of backbone design was analyzed across convolutional and transformer-based architectures. Since the stage-two classification backbones were evaluated using the same RGB + ExG + Hue five-channel input representation, the observed performance differences among these classifiers mainly reflect the effect of backbone design rather than input modality.

As shown in Table 8, CP2-A2C2F achieved the highest accuracy, weighted F1 score, and the lowest MAE among all evaluated models. These findings suggest that the proposed

backbone structure is effective for preserving localized chromatic information required for fine-grained strawberry ripeness classification under the tested greenhouse conditions.

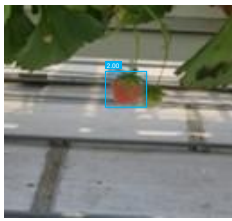
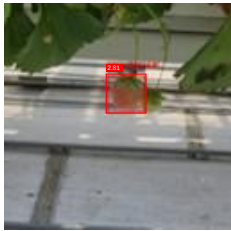
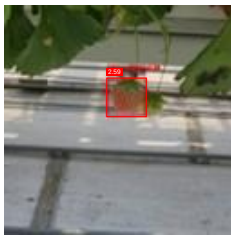
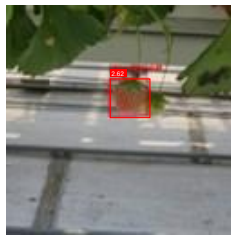
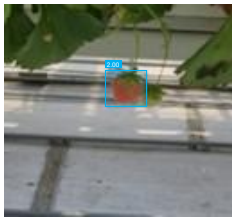
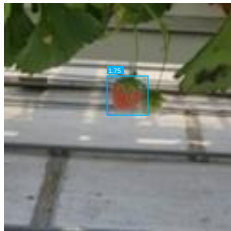
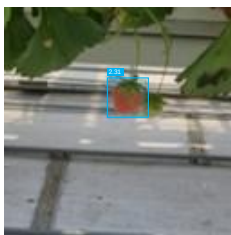
GT	ResNet50	ConvNeXtV2	MobileNetV3
			
Class: Turning Reg.: 2.00	Class: Late turning Reg.: 2.81	Class: Late turning Reg.: 2.59	Class: Late turning Reg.: 2.62
SwinV2	A2C2F	Ours	
			
Class: Turning Reg.: 2.00	Class: Turning Reg.: 1.75	Class: Turning Reg.: 2.31	Class: Turning Reg.: 2.01

Figure 7. Qualitative comparison of strawberry ripeness prediction results across different backbone models. The regression score follows the ordinal ripeness scale from 0.0 to 4.0, and the class shown below each image is interpreted from the nearest ordinal ripeness level. Blue boxes indicate ROIs for which the ripeness class was correctly predicted, whereas red boxes indicate ROIs for which the ripeness class was incorrectly predicted.

Transformer-based architectures such as SwinV2 can capture global contextual relationships, while convolution-based models are effective for extracting local spatial patterns. However, strawberry ripeness classification requires both localized color transition information and broader maturity context. This is particularly important for intermediate ripeness classes, where red pigmentation appears only in limited regions of the fruit surface.

The CP2-A2C2F backbone was designed to address this requirement by preserving high-resolution P2 features and integrating them with deeper P4 semantic representations through the CP2 Fusion module. This structure enables the model to maintain fine scale color and texture information while utilizing high level contextual features. This design is also consistent with the backbone structure shown in Figure 3, where P2 and P4 features are directly connected to the final fusion stage.

Overall, the results indicate that combining the RGB + ExG + Hue five-channel representation with the CP2-A2C2F backbone provides a useful strategy for modeling continuous strawberry maturation patterns, showing moderately improved overall performance and lower ordinal MAE compared with the evaluated backbone architectures within this dataset.

3.4. Ablation Study

An ablation study was conducted to evaluate the effectiveness of the proposed design components. All ablation experiments in this section were evaluated on the validation set, which was used for model selection and configuration analysis.

As summarized in Table 12, the input-channel ablation was revised to focus on two representative backbones, A2C2F and the proposed CP2-A2C2F. For A2C2F, adding ExG

and Hue resulted in only marginal changes in accuracy and weighted F1 score, while slightly reducing MAE from 0.188 to 0.179. This indicates that the additional color-derived channels had a limited effect on overall classification performance, although they may have contributed to a small reduction in ordinal prediction error.

For CP2-A2C2F, adding ExG alone did not improve performance over RGB, as accuracy decreased from 0.875 to 0.864 and MAE increased from 0.171 to 0.210. In contrast, when ExG and Hue were used together, the RGB + ExG + Hue configuration showed the most favorable balance among the evaluated input representations, with an accuracy of 0.881, a weighted F1 score of 0.883, and an MAE of 0.168. Although the improvement over RGB alone was moderate, this result suggests that the proposed CP2-A2C2F backbone appears to utilize the combined five-channel representation more effectively than the RGB+ExG configuration. Therefore, the benefit of the proposed input representation should be interpreted mainly as a configuration-dependent effect of using ExG and Hue together within the CP2-A2C2F backbone, rather than as evidence that each additional color-derived channel independently improves performance.

Table 12. Input-channel ablation study for strawberry ripeness classification, evaluated on the validation set.

Model	Input	Accuracy	Weighted F1	MAE ↓
A2C2F	RGB	0.877	0.879	0.188
	RGB + ExG	0.877	0.881	0.186
	RGB + ExG + Hue	0.878	0.881	0.179
Ours (CP2-A2C2F)	RGB	0.875	0.877	0.171
	RGB + ExG	0.864	0.868	0.210
	RGB + ExG + Hue	0.881	0.883	0.168

Note: Bold values indicate the best performance among the evaluated input configurations for each model and metric.

Beyond the input-channel analysis, the impact of different feature fusion strategies within the CP2-A2C2F backbone was further evaluated on the validation set. To examine the contribution of the intermediate P3 feature map, four fusion configurations were compared: P2 + P3, P3 + P4, P2 + P3 + P4, and the proposed P2 + P4 fusion.

As summarized in Table 13, the proposed P2 + P4 fusion achieved the highest validation accuracy of 0.881, the highest weighted F1 score of 0.883, and the lowest MAE of 0.168 among the evaluated fusion strategies. In comparison, directly including the intermediate P3 feature map did not further improve the validation performance. These results suggest that the direct fusion of high-resolution P2 features and deeper P4 semantic features provides a more effective balance between class discrimination and ordinal ripeness estimation than the other evaluated fusion configurations.

Therefore, the P2 + P4 configuration was selected as the final fusion strategy.

To examine the sensitivity of the proposed framework to the regression loss weighting coefficient λ_{reg} , an additional ablation study was conducted across five candidate values on the validation set.

Table 13. Feature fusion ablation study of the CP2-A2C2F backbone using the RGB + ExG + Hue representation, evaluated on the validation set.

Variant	Accuracy	Weighted F1	MAE ↓
P2 + P3	0.863	0.865	0.190
P3 + P4	0.865	0.865	0.181
P2 + P3 + P4	0.872	0.872	0.184
Ours (P2 + P4)	0.881	0.883	0.168

Note: Bold values indicate the best performance among the evaluated feature fusion configurations for each metric.

As shown in Table 14 and Figure 8, the MAE values in this analysis were recomputed as the ordinal distance between the class predicted by the classification head and the ground-truth ripeness class. The raw output of the auxiliary regression head was not used because, when $\lambda_{\text{reg}} = 0.00$, the regression head receives no gradient from the training objective and therefore does not provide a meaningful ordinal prediction. This classification-based definition allows all values of λ_{reg} to be compared using the same measure.

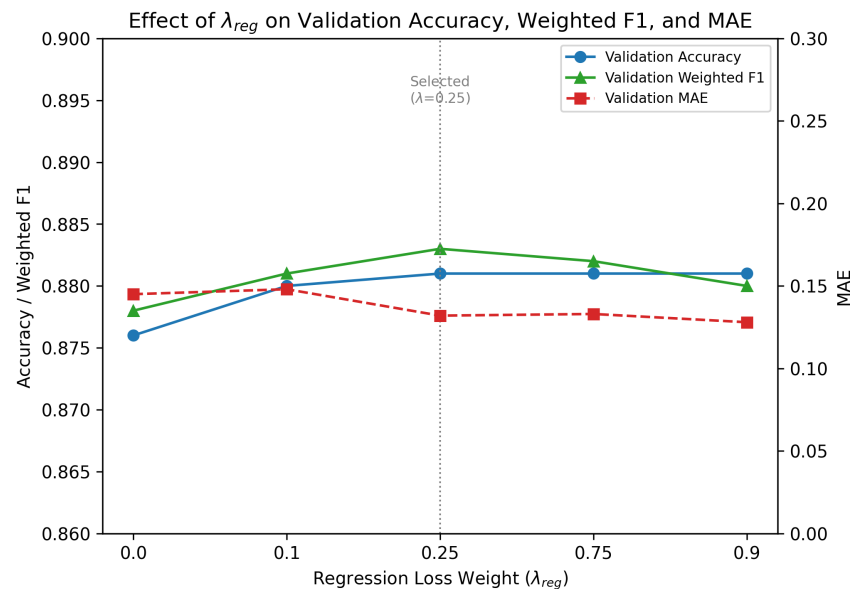


Figure 8. Validation accuracy, weighted F1, and classification-based ordinal MAE for different values of λ_{reg} . The MAE was computed from the ordinal distance between the classification head’s predicted class and the ground-truth class.

Table 14. Effect of the regression loss weighting coefficient λ_{reg} on validation performance of the CP2-A2C2F model.

λ_{reg}	Accuracy	Weighted F1	MAE ↓
0.00	0.876	0.878	0.145
0.10	0.880	0.881	0.148
0.25	0.881	0.883	0.132
0.75	0.881	0.882	0.133
0.90	0.881	0.880	0.128

Note: Bold values indicate the best performance for each metric among the evaluated λ_{reg} settings.

The MAE did not decrease monotonically as λ_{reg} increased. The setting $\lambda_{\text{reg}} = 0.10$ produced a slightly higher MAE than the classification-only setting (0.148 versus 0.145), whereas $\lambda_{\text{reg}} = 0.25, 0.75$, and 0.90 produced lower MAE values ranging from 0.128 to 0.133. Classification performance varied only slightly for $\lambda_{\text{reg}} \geq 0.10$, with accuracy ranging from 0.880 to 0.881 and weighted F1 ranging from 0.880 to 0.883. The setting $\lambda_{\text{reg}} = 0.25$ was retained because it achieved the joint highest accuracy of 0.881, the highest weighted F1 score of 0.883, and a low classification-based ordinal MAE of 0.132. Although $\lambda_{\text{reg}} = 0.90$ produced the numerically lowest MAE of 0.128, its weighted F1 score was lower at 0.880. Therefore, $\lambda_{\text{reg}} = 0.25$ provided the most favorable balance between classification performance and ordinal prediction error among the evaluated settings.

3.5. Computational Efficiency and Inference Latency

Since the proposed framework is intended for harvest decision support, the computational cost of the two-stage pipeline was also evaluated. Table 15 reports the parameter count, FLOPs, and per-stage latency measured on an NVIDIA RTX 3080 GPU with a batch

size of 1. The YOLO11m detector accounted for most of the parameters (20.05 M) and FLOPs (153.42 GFLOPs), reflecting its 960×960 input resolution, while CP2-A2C2F was substantially lighter (2.67 M parameters, 6.67 GFLOPs). Detection required 16.06 ms per image, and ROI cropping and five-channel construction added 1.26 ms and 7.36 ms, respectively. Classification of the detected ROIs required 12.75 ms on average. The complete pipeline reached an end-to-end latency of 37.42 ms per image, corresponding to a throughput of approximately 26.7 images per second, with a peak GPU memory usage of 390.1 MB. These results indicate that the proposed framework can operate at a processing speed suitable for practical greenhouse deployment on a single consumer grade GPU.

Table 15. Parameter count, FLOPs, and per-stage inference latency of the two-stage pipeline (NVIDIA RTX 3080, batch size 1).

Stage	Params (M)	GFLOPs	Latency (ms)
Detection (YOLO11m)	20.05	153.42	16.06 ± 1.94
ROI Cropping	—	—	1.26 ± 0.69
Five-channel Construction	—	—	7.36 ± 3.45
Classification (CP2-A2C2F)	2.67	6.67	12.75 ± 8.33
Total (end-to-end)	22.73	160.10	37.42 ± 10.38

4. Conclusions

In this study, we presented a two-stage vision-based framework for five-class strawberry ripeness grading to support harvest decision making in smart greenhouse environments. The framework first localizes strawberry regions using a YOLO11m detector and then classifies the ripeness class of each extracted ROI using the CP2-A2C2F backbone. This design addresses the difficulty of distinguishing gradual color transitions that occur among intermediate ripeness classes under the tested greenhouse conditions.

The main contribution of this study is the combination of the RGB + ExG + Hue five-channel representation and the CP2-A2C2F backbone architecture. By adding Excess Green (ExG) and Hue channels to the conventional RGB input, the model can utilize additional chromatic information related to green dominant early classes and gradual red coloration during maturation. In addition, the CP2-A2C2F backbone preserves high-resolution P2 features and integrates them with deeper P4 semantic representations through a dual scale fusion strategy. This structure helps maintain localized color and texture information while also using high level contextual features for ripeness grading.

The proposed framework was evaluated within a single commercial greenhouse facility under realistic cultivation conditions. The CP2-A2C2F model achieved the lowest MAE of 0.162 on the ordinal ripeness scale, reflecting the smallest average ordinal deviation between predicted and actual ripeness stages, together with the best overall accuracy of 87.7% and a weighted F1 score of 0.877. In the input channel ablation study, the RGB + ExG + Hue configuration reduced the MAE of CP2-A2C2F from 0.171 to 0.168, corresponding to an approximately 1.8% reduction in ordinal prediction error, while also improving accuracy from 87.5% to 88.1% compared with the RGB-only configuration. The repeated-seed confusion matrix analysis further indicated that the class-wise improvement was most evident in the early turning class, where subtle red pigmentation first appears, while adjacent-stage ambiguity remained for the turning and late turning classes. However, the class-wise improvement was not uniform. In particular, the F1 score for the Turning class decreased from 0.680 with A2C2F to 0.605 with CP2-A2C2F, despite the improvements in overall accuracy, weighted F1, and ordinal MAE, indicating that the most ambiguous intermediate stage remains challenging.

These findings indicate that combining explicit color-based input representation with P2 and P4 feature fusion is effective for modeling continuous strawberry maturation patterns within the present dataset. Since the framework uses standard RGB images and derived color channels rather than expensive multispectral hardware, it offers a non-destructive approach for automated strawberry ripeness assessment within the tested greenhouse environment.

Future work will focus on real-time edge computing deployment for autonomous harvesting systems. Field validation studies comparing model driven ripeness grading with manual grading performed by experienced workers will also be conducted to assess the practical utility of the proposed framework for on site harvest decision making. A condition-stratified evaluation across illumination and occlusion levels was not conducted in the present study; therefore, performance under these specific conditions remains unverified. The dataset will be expanded to include multiple strawberry cultivars, diverse greenhouse facilities, different imaging devices, broader seasonal conditions, and controlled lighting or degraded vision conditions, such as illumination changes, partial occlusion, motion blur, and backlight, to further evaluate the generalizability, robustness, and downstream performance of the proposed approach under more varied greenhouse environments.

Author Contributions: Conceptualization, G.L. and B.K.; methodology, G.L. and B.K.; software, G.L.; validation, S.C. and B.K.; formal analysis, G.L.; investigation, G.L. and D.G.; resources, S.J.; data curation, G.L., B.K., J.S. and S.C.; writing—original draft preparation, G.L.; writing—review and editing, G.L., D.G., S.C., J.S., S.J. and B.K.; visualization, G.L., S.C. and D.G.; supervision, S.J. and B.K.; project administration, S.J.; funding acquisition, B.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) and the Korea Smart Farm R&D Foundation (KosFarm) through the Smart Farm Innovation Technology Development Program, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA), the Ministry of Science and ICT (MSIT), and the Rural Development Administration (RDA) (Project No. RS-2025-02312929).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Skrovankova, S.; Sumczynski, D.; Mlcek, J.; Jurikova, T.; Sochor, J. Bioactive compounds and antioxidant activity in different types of berries. *Int. J. Mol. Sci.* **2015**, *16*, 24673–24706. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Kishimoto, Y.; Taguchi, C.; Iwashima, T.; Kobayashi, T.; Kikoku, Y.; Nishiyama, H.; Masuda, Y.; Kondo, K. Effects of acute strawberry consumption on serum levels of vitamin C and folic acid, the antioxidant potential of LDL and blood glucose response: A randomised cross-over controlled trial. *J. Nutr. Sci.* **2023**, *12*, e39. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Muthukumaran, S.; Tranchant, C.; Shi, J.; Ye, X.; Xue, S.J. Ellagic acid in strawberry (*Fragaria* spp.): Biological, technological, stability, and human health aspects. *Food Qual. Saf.* **2017**, *1*, 227–252. [\[CrossRef\]](#)
4. Tang, C.; Chen, D.; Wang, X.; Ni, X.; Liu, Y.; Liu, Y.; Mao, X.; Wang, S. A fine recognition method of strawberry ripeness combining Mask R-CNN and region segmentation. *Front. Plant Sci.* **2023**, *14*, 1211830. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Saeed, M.T.; Anwar, R.; Khan, A.S.; Ziaf, K.; Hussain, Z.; Razi, F.U.D.; Hasni, G.M.; Mariyum, A. Impact of Maturity Stage on Postharvest Quality of Strawberries Cold Stored for Fresh Consumption. *Plant Bull.* **2023**, *2*, 142–150. [\[CrossRef\]](#)
6. Pineli, L.d.L.d.O.; Moretti, C.L.; dos Santos, M.S.; Campos, A.B.; Brasileiro, A.V.; Córdova, A.C.; Chiarello, M.D. Antioxidants and other chemical and physical characteristics of two strawberry cultivars at different ripeness stages. *J. Food Compos. Anal.* **2011**, *24*, 11–16. [\[CrossRef\]](#)

7. Xu, L.; Zhao, Y. Automated strawberry grading system based on image processing. *Comput. Electron. Agric.* **2010**, *71*, S32–S39. [CrossRef]
8. Oo, L.M.; Aung, N.Z. A simple and efficient method for automatic strawberry shape and size estimation and classification. *Biosyst. Eng.* **2018**, *170*, 96–107. [CrossRef]
9. Cho, W.; Na, M.; Kim, S.; Jeon, W. Automatic prediction of brix and acidity in stages of ripeness of strawberries using image processing techniques. In Proceedings of the 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Republic of Korea, 23–26 June 2019; pp. 1–4.
10. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
11. Zhao, S.; Fang, C.; Hua, T.; Jiang, Y. Detecting the Maturity of Red Strawberries Using Improved YOLOv8s Model. *Agriculture* **2025**, *15*, 2263. [CrossRef]
12. Chen, Y.; Xu, H.; Chang, P.; Huang, Y.; Zhong, F.; Jia, Q.; Chen, L.; Zhong, H.; Liu, S. CES-YOLOv8: Strawberry maturity detection based on the improved YOLOv8. *Agronomy* **2024**, *14*, 1353. [CrossRef]
13. Liu, J.; Guo, J.; Zhang, S. YOLOv11-HRS: An improved model for strawberry ripeness detection. *Agronomy* **2025**, *15*, 1026. [CrossRef]
14. Yu, H.; Qian, C.; Chen, Z.; Chen, J.; Zhao, Y. Ripe-detection: A lightweight method for strawberry ripeness detection. *Agronomy* **2025**, *15*, 1645. [CrossRef]
15. Zheng, C.; Abd-Elrahman, A.; Whitaker, V. Remote sensing and machine learning in crop phenotyping and management, with an emphasis on applications in strawberry farming. *Remote Sens.* **2021**, *13*, 531. [CrossRef]
16. Rammal, A.; Assaf, R.; Perrin, E.; Vrabie, V. Enhancing strawberry maturity assessment using mid-infrared spectral analysis with advanced variable selection and supervised classification. *Sci. Rep.* **2026**, *16*, 10154. [CrossRef] [PubMed]
17. Feng, Y.; Wang, Y.; Purohit, A.; Liang, J.; Wang, X.; Qiao, M.; Xiao, Z.; Luo, Y. Classification of Strawberry Maturity Using Swin Transformer V2: Integrating Near-Infrared Spectral, Chemical, and Optical Properties. *J. Future Foods* **2025**, *7*, 81–89. [CrossRef]
18. Yee-Rendon, A.; Torres-Pacheco, I.; Trujillo-Lopez, A.S.; Romero-Bringas, K.P.; Millan-Almaraz, J.R. Analysis of New RGB Vegetation Indices for PHYVV and TMV Identification in Jalapeño Pepper (*Capsicum annuum*) Leaves Using CNNs-Based Model. *Plants* **2021**, *10*, 1977. [CrossRef] [PubMed]
19. Qian, Y.; Xiao, Z.; Deng, Z. Fine-grained crop pest classification based on multi-scale feature fusion and mixed attention mechanisms. *Front. Plant Sci.* **2025**, *16*, 1500571. [CrossRef] [PubMed]
20. Li, Y.; Li, C.; Zhu, T.; Zhang, S.; Liu, L.; Guan, Z. A recognition model for winter peach fruits based on improved ResNet and multi-scale feature fusion. *Front. Plant Sci.* **2025**, *16*, 1545216. [CrossRef] [PubMed]
21. Qu, X.; Ma, L.; Shen, T.; Han, X.; Yan, X.; Bi, S. Ordinal regression-extreme learning machine based apple grading method. In Proceedings of the 2022 China Automation Congress (CAC), Xiamen, China, 25–27 November 2022; pp. 3507–3511.
22. Nguyen, B.A.; Kha, M.B.; Dao, D.M.; Nguyen, H.K.; Nguyen, M.D.; Nguyen, T.V.; Rathnayake, N.; Hoshino, Y.; Dang, T.L. UFR-GAN: A lightweight multi-degradation image restoration model. *Pattern Recognit. Lett.* **2025**, *197*, 282–287. [CrossRef]
23. Jocher, G.; Qiu, J. Ultralytics YOLO11. Version 11.0.0. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 2 July 2026).
24. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 78433–78457.
25. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
26. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
27. Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8430–8439.
28. Korea Rural Economic Institute. Agricultural Outlook Monthly Report. 2024. Available online: <https://aglook.krei.re.kr/main/uObserveMonth/OVR0000000031?queryType=2024&queryType2=7811> (accessed on 1 August 2024).
29. Sánchez-Sevilla, J.F.; Botella, M.A.; Valpuesta, V.; Sanchez-Vera, V. Autophagy is required for strawberry fruit ripening. *Front. Plant Sci.* **2021**, *12*, 688481. [CrossRef] [PubMed]
30. Elhariri, E.; El-Bendary, N.; Saleh, S.M. Strawberry-DS: Dataset of annotated strawberry fruits images with various developmental stages. *Data Brief* **2023**, *48*, 109165. [CrossRef] [PubMed]
31. Woebbecke, D.M.; Meyer, G.E.; Von Bargen, K.; Mortensen, D.A. Color indices for weed identification under various soil, residue, and lighting conditions. *Trans. ASAE* **1995**, *38*, 259–269. [CrossRef]
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

33. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
35. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
36. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
37. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
38. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
39. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
42. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
43. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.