Journal of Zhejiang University-SCIENCE B 2026 27(5):466-481
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.B2500647

Check for updates

Embedding of ripening topology into one-stage detection for tomato cluster phenotyping

Bingquan CHU[1], Ruiyuan WU[1], Haijun ZHANG[1], Haochuan QIN[1], Zishun PENG[1], Fengle ZHU[2]✉, Yong HE[3]✉

[1]*School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China*
[2]*College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou 310023, China*
[3]*College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China*

Abstract: The automated assessment of tomato ripeness is vital for modern greenhouse operations, yet challenges remain due to variable environmental conditions. To provide a solution, we propose rank-aware You Only Look Once (YOLO), a novel detection framework that incorporates the biological prior of top-to-bottom ripening within fruit clusters. This is achieved through two key innovations: an efficient position-aware head for regressing relative height for fruits and a dynamic margin-aware ranking loss (DM-RankLoss) that enforces the correct spatial sequence. Evaluated on a 3500-image dataset from a solar greenhouse, our plug-and-play module could boost the mean average precision (mAP) at intersection over union (IoU) threshold of 0.50 (mAP_{50}) of multiple YOLO architectures by up to 5.66 pecentage points. The model effectively learns the cluster topology, achieving a height-mean absolute error (H-MAE) of 0.107 (normalized) and a pairwise ranking accuracy (PRA) of 84.59%, while it reduces the parameter count by over 10% compared to the baseline for efficient deployment. Visualizations confirm that the model leverages spatial context to resolve color ambiguities. Our work offers a sensor-free, accurate, and efficient solution for in situ phenotyping in agricultural robotics.

Key words: Tomato ripeness; Phenotype; Object detection; Topology; You Only Look Once (YOLO); Spatial sequence

1 Introduction

Precision agriculture represents a critical technological direction for addressing global food security challenges and improving agricultural productivity. Tomato is one of the most important vegetable crops worldwide, with its production operating on a massive scale. According to the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), global tomato production reached approximately 186.1 million tons in 2022, with China contributing nearly 37% of the global total (Food and Agriculture Organization of the United Nations (FAO), 2023). However, in contrast to the highly automated greenhouse systems used in countries such as the Netherlands, tomato harvesting in major producing regions like China remains predominantly dependent on manual ripeness assessment. This approach is not only inefficient and labor-intensive but also susceptible to subjective judgment, variable lighting conditions, and tomato occlusion. These limitations often lead to inconsistent product quality, thereby undermining the standardization and economic sustainability of the supply chain (Wang et al., 2022). Therefore, the central challenge in this field extends beyond improving efficiency to resolving the quality control and standardization issues engendered by the inherent subjectivity of manual inspection.

Computer vision technology, particularly deep learning-based object detection, has been increasingly applied to agricultural phenotyping (Yao et al., 2025). Single-stage detectors, represented by the You Only Look Once (YOLO) series (Albahar, 2023), have emerged as mainstream solutions for fruit and vegetable recognition due to their real-time performance and end-to-end capability. Li RZ et al. (2023) adapted YOLOv5s to classify four tomato ripening stages; Chen

✉ Fengle ZHU, zhufl@zjut.edu.cn
Yong HE, yhe@zju.edu.cn
Fengle ZHU, https://orcid.org/0009-0000-5610-4448
Yong HE, https://orcid.org/0000-0001-6752-1757
Bingquan CHU, https://orcid.org/0009-0009-2319-4454

Received Oct. 14, 2025; Revision accepted Mar. 12, 2026; Crosschecked Apr. 8, 2026

WJ et al. (2024) integrated multi-modal data (RGB, depth, and near-infrared) with YOLOv7 in a model named YOLO-DNA for ripeness detection; and another study by Chen WB et al. (2024) employed a multi-decoder multi-task detection YOLOv7 (MTD-YOLOv7) architecture to simultaneously identify tomatoes and clusters along with their maturity levels. These methods have all achieved high mean average precision (mAP) on dedicated datasets.

Despite certain advances, a common limitation persists: current approaches treat ripeness primarily as a visual attribute determined solely by the color, texture, or multi-spectral characteristics of individual tomatoes (Khan et al., 2025). This over-reliance on apparent features renders models vulnerable to complex agricultural environments such as lighting variations, shadows, and occlusions—problems reminiscent of those faced by manual assessment (Xiao et al., 2023). Although some researchers have attempted to enhance robustness through diverse data collection or the incorporation of additional sensors (e.g., infrared or depth cameras), these strategies essentially rely on data quantity or hardware enhancements rather than addressing the core issue through improved model reasoning.

The above issues highlight a critical research gap: existing methods overlook the deterministic physical-topological prior knowledge inherent in tomato cluster structures. Specifically, they fail to utilize a crucial biological principle that tomatoes on the same cluster ripen in a deterministic spatial sequence, progressing from the proximal (stem-attached) to the distal end (Gautier et al., 2005). As such, the ripening process of on-cluster tomatoes follows a top-to-bottom sequence on the cluster, illustrating the transition from the green (immature) stage to the fully ripe (mature) stage, as shown in Fig. 1. Such structural information provides an invaluable clue for determining ripeness, particularly in visually ambiguous scenarios where lighting variations obscure color information or where tomatoes at different stages exhibit similar appearances.

This study proposes a fundamental reformulation of the tomato ripeness detection problem. Rather than treating it as a conventional multi-class object detection task (i.e., independently classifying ripeness into one of four discrete and ostensibly unrelated categories: “green,” “breaking,” “ripe,” or “fully ripe”), it should be defined as a joint task of spatially-aware sequence regression. The rationale for this reformulation lies in the inherent ordinal nature of ripening: the four stages represent a continuous developmental sequence rather than arbitrary labels. This makes tomato ripeness prediction a suitable application for ordinal regression, a fundamental area of machine learning dedicated to modeling ordered classes, which leverages the structure between labels to improve both accuracy and interpretability (Albahar, 2023). Crucially, the spatial layout of tomatoes within a cluster provides a “physical anchor” for this ordinal relationship. Thus, predicting the ripeness of a tomato is equivalent to estimating its position within the “spatio-temporal developmental sequence” of its cluster. This approach aligns with established findings in other computer vision domains, such as monocular depth estimation, age estimation, and ranking, where ordinal regression has been shown to enhance robustness by explicitly modeling the relationships between adjacent labels, thereby reducing the impact of label noise and ambiguity (Fu et al., 2018; Díaz and Marathe, 2019; Cao et al., 2020).

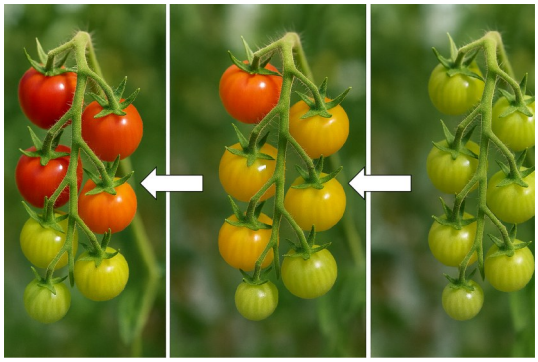

Fig. 1 Ripening process of on-cluster tomatoes, illustrating the transition from the green (immature) stage to the fully ripe (mature) stage. This progression follows a top-to-bottom sequence on the cluster.

Shifting the ripeness determination perspective in this manner transforms the core task from asking “To which class does this tomato belong?” to “Where does this tomato fall within its cluster’s developmental sequence?” While the former is prone to visual ambiguity (for example, under challenging lighting conditions, the colors of “ripe” and “fully ripe” tomatoes may appear indistinguishable due to inter-class similarity), the latter leverages structural regularities, such as the vertical relative height (h_{rel}) of a tomato within a cluster, as a robust anchor for inference. By transforming a visually confusable classification problem into a structurally grounded ordinal regression one, models can reason more reliably when visual cues are compromised.

To implement this reformulated approach, we propose the rank-aware YOLO framework. This architecture concretely embodies the new paradigm through two core innovations: (1) an efficient position-aware head, which is a lightweight extension to a standard YOLO detection head that incorporates a regression branch for predicting the h_{rel} of each tomato, thereby enabling spatial-ordinal regression; and (2) a dynamic margin-aware ranking loss (DM-RankLoss), a novel domain-specific loss function designed to supervise the joint detection-and-ranking task. Unlike generic losses, DM-RankLoss incorporates biological prior knowledge by explicitly penalizing predictions that violate the natural ripening sequence, thus guiding the optimization process with structured domain constraints.

In summary, the main contributions of this paper are as follows:

(1) On-cluster tomato ripeness detection is redefined as a unified task combining object detection and spatially-aware sequence regression. This paradigm offers a more robust and principled solution to overcome the limitations of conventional methods that rely excessively on ambiguous visual features.

(2) The rank-aware YOLO framework is introduced, which integrates a lightweight position-aware prediction head and a novel DM-RankLoss function that explicitly encodes biological ripening priors into the learning objective.

(3) Extensive experiments on a real-world greenhouse dataset demonstrate that our method significantly improves detection accuracy and generalization ability across multiple YOLO architectures, without introducing additional hardware cost or significant computational overhead.

2 Materials and methods

2.1 Data acquisition

2.1.1 Dataset sample collection

The on-cluster tomato image dataset used in this study was collected from a solar greenhouse located at the Wuzhen International Internet Agriculture Expo Park in Tongxiang, Zhejiang Province, China. Data acquisition spanned 12 d between March 15 and May 20, 2025, covering multiple growth stages of the tomato plants. Image collection was performed on the following dates: March 15, 18, 22, and 28; April 5, 12, 18, and 25; May 2, 8, 15, and 20.

Image acquisition was conducted using an OAK-D Pro Wide stereo depth camera (Luxonis, Littleton, CO, USA) mounted on a self-developed farm patrol robot. The robot followed a predefined path for automated navigation, with the camera fixed to an adjustable pole to control capture height. Video recordings were captured at 1920×1080 resolution and 30 frames per second (FPS) from multiple viewpoints, specifically at vertical heights of 1.2, 1.5, and 1.8 m above ground level, and horizontal distances of 0.3–0.8 m from the target plants. The image acquisition system consisted of a farm patrol robot equipped with a camera that captures images from multiple heights and distances within the solar greenhouse environment (Fig. 2).

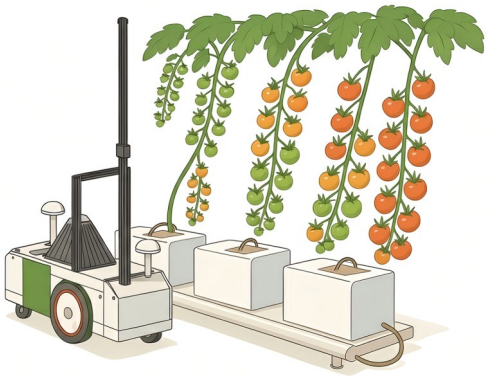

Fig. 2 Schematic of the image acquisition system for on-cluster tomatoes in a solar greenhouse. A farm patrol robot equipped with an OAK-D Pro Wide camera captures images from multiple heights and distances.

After randomly extracting frames from the collected videos and discarding excessively blurry or abnormally exposed images, a final dataset of 3500 images was assembled. The dataset was then randomly divided into training, validation, and test subsets at a ratio of 7:2:1, resulting in 2450 images for training, 700 for validation, and 350 for testing. To support subsequent style transfer tasks, a supplementary set of images was also collected under varying lighting conditions, including morning, dusk, and artificial night lighting, for training the style transfer model.

2.1.2 Dataset annotation rules and semi-automatic annotation process

Tomato fruit ripeness was classified into four stages: green, breaking, ripe, and fully ripe, based on the

National Standard GH/T 1193-2021 "Tomato" (All China Federation of Supply and Marketing Cooperatives, 2021) and its harvesting requirements (Li XX et al., 2023), as follows:

(1) Green: fruit reaches full size, with a glossy surface and whitish-green skin.

(2) Breaking: yellowish-green or light-red streaks appear around the blossom end, and the entire fruit may turn pale yellow in later phases.

(3) Ripe: fruit displays a relatively uniform orange hue, with red coloration covering less than 30% of the surface, primarily around the blossom end.

(4) Fully ripe: fruit is distinctly red, with more than 40% of the surface showing red coloration.

Cluster ripeness was classified into two levels: mature and immature. A cluster was considered mature if the bottom two fruits had reached at least the breaking stage; otherwise, it was classified as immature (Chen WB et al., 2024).

A semi-automatic annotation process was adopted to label the images. Initially, 500 images were manually annotated using the Roboflow platform. The remaining 3000 images were then pre-annotated via the platform's "auto label" feature, followed by comprehensive manual verification to ensure label quality and correct errors. This hybrid approach improves annotation efficiency while reducing the subjectivity inherent to fully manual labeling.

Upon the completion of ripeness annotation, an automated script was employed to annotate each fruit with two additional attributes: clusterID and h_{rel} values. A schematic diagram of this extended labeling format is presented, together with a visualization of the assigned attributes on a representative sample image in Fig. 3. Each detected cluster was assigned a unique identifier (ID), numbered sequentially from top to bottom. Based on their two-dimensional (2D) spatial coordinates (x, y), fruits were associated with a specific cluster if their center point fell within its bounding box. The script then calculated the normalized h_{rel} of each fruit within its cluster. The resulting annotations were saved in .txt format for training our modified YOLO model.

To quantify the topological position of each fruit within its cluster, we define h_{rel} as a normalized metric derived from the 2D spatial coordinates. Let $B_i=\{x_i, y_i, w_i, h_i\}$ denote the bounding box of the ith tomato fruit belonging to a specific cluster C, where x_i and y_i are the normalized coordinates of the fruit's geometric center

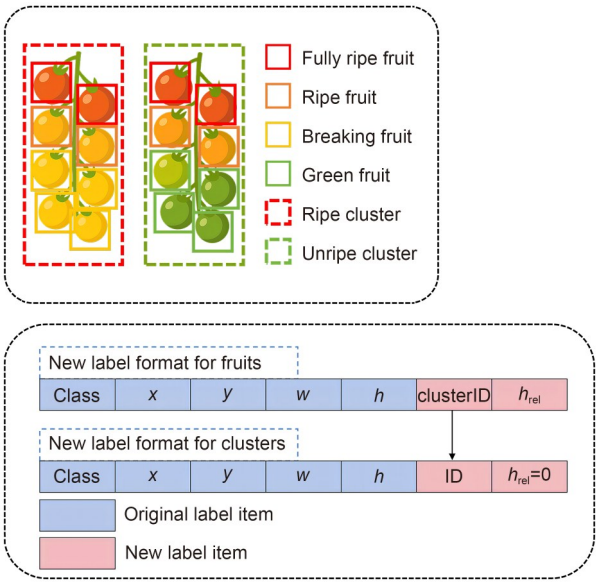

Fig. 3 Extended YOLO label format and dataset annotation. The standard five-parameter format (class, *x*, *y*, *w*, and *h*) is enhanced with two attributes: clusterID and h_{rel}. Class is the ripeness category; *x* and *y* are the normalized center coordinates; *w* and *h* are the normalized width and height, respectively; clusterID is the unique identifier (ID) of the cluster; and h_{rel} is the normalized relative height within the cluster.

and w_i and h_i are the normalized width and height, respectively. These values are normalized by the width (W) and height (H) of the input image ($x_i=x_{pixel}/W$, $y_i=y_{pixel}/H$, $w_i=w_{pixel}/W$, $h_i=h_{pixel}/H$). The vertical extent of cluster C is defined by its uppermost boundary y^C_{top} and lowermost boundary y^C_{bottom}, calculated as the extremum of all constituent fruits:

$$y^C_{top}=\min\left(y_i-\frac{h_i}{2}\right),\ y^C_{bottom}=\max\left(y_i+\frac{h_i}{2}\right). \quad (1)$$

The normalized relative height for fruit i (h^i_{rel}) is formulated as:

$$h^i_{rel}=\frac{y_i-y^C_{top}}{y^C_{bottom}-y^C_{top}+\varepsilon}, \quad (2)$$

where ε is a small constant (1×10^{-6}) ensuring numerical stability. This normalization maps the vertical position of the fruit to a continuous range, where 0 corresponds to the top end and 1 corresponds to the bottom end of the cluster.

Detailed information on the style transfer-based data augmentation is provided in Method S1.

2.2 Rank-aware YOLO framework

2.2.1 Framework overall

For detailed background on the YOLOv12 architecture, please refer to Method S2.

To effectively leverage the critical physical-spatial prior knowledge that on-cluster tomatoes ripen from top to bottom (Gautier et al., 2005), we propose a rank-aware YOLO framework, whose overall architecture is illustrated in Fig. 4. The core of this framework is a plug-and-play "rank-aware module" integrated into a standard YOLO detector, comprising an efficient position-aware head and a novel DM-RankLoss. The data processing pipeline consists of the following stages:

(1) Feature extraction and fusion: The input image is processed through an enhanced YOLOv12 backbone incorporating multi-dimensional collaborative attention (MCAM) to extract multi-scale features dynamically. These features are effectively fused in the neck network using a path aggregation network (PAN) structure, which builds upon the feature pyramid network (FPN) with bottom-up pathways to improve localization through multi-scale feature aggregation. To ensure that performance improvements are attributable solely to our proposed module, the backbone and neck architectures of the baseline model remain unmodified.

(2) Position-aware multi-task prediction: The fused features are passed into our custom efficient

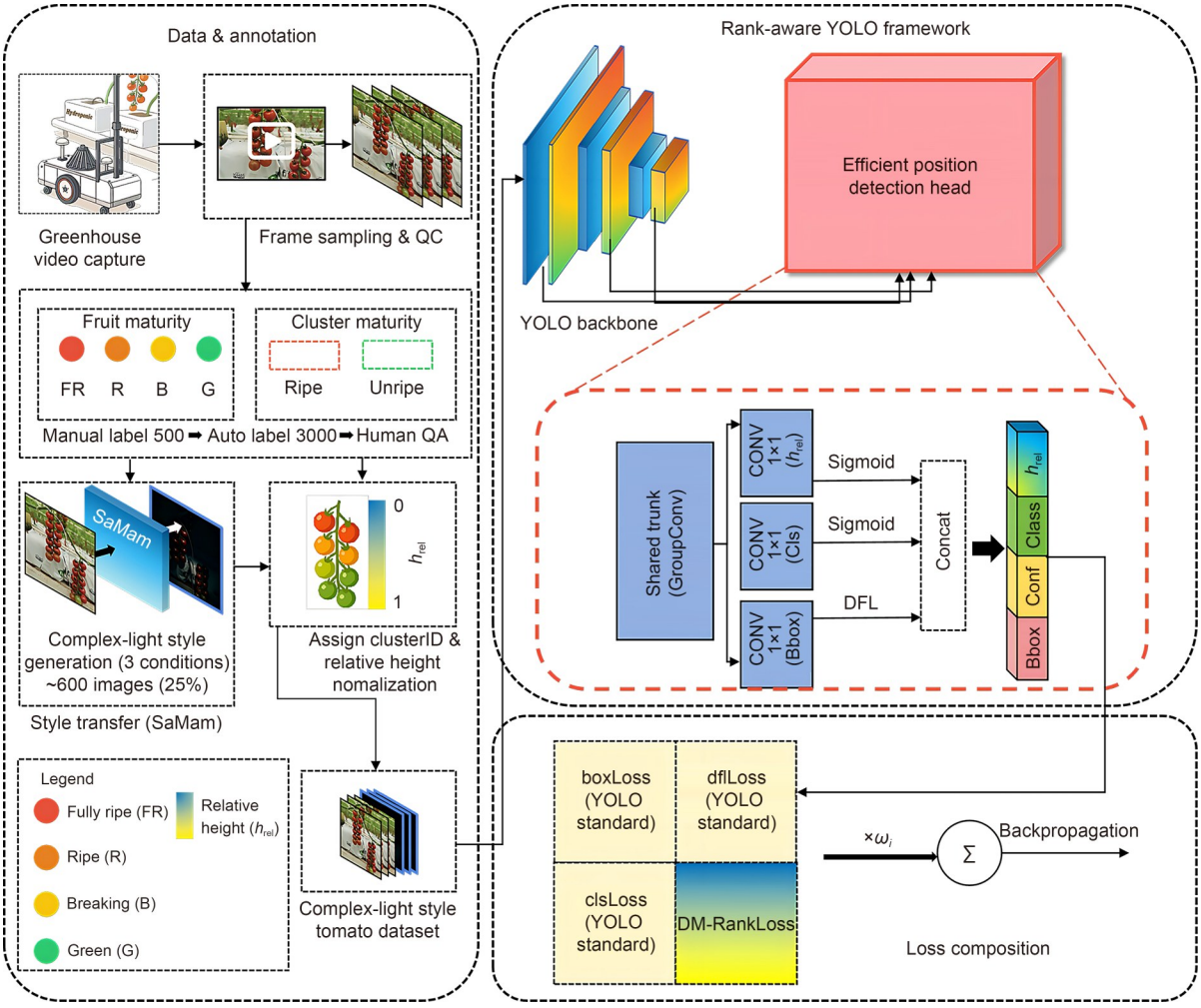

Fig. 4 Overview of the proposed rank-aware YOLO framework. The pipeline comprises data acquisition, style transfer-based augmentation, feature extraction, and multi-task prediction via the efficient position-aware head, supervised by a composite loss function incorporating dynamic margin-aware ranking loss (DM-RankLoss). QC: quality control; QA: quality assurance; GroupConv: grouped convolution; CONV: convolutional layer; Bbox: bounding box; Cls: classification; DFL: distribution focal loss; Conf: confidence score representing the probability of object existence and localization accuracy; ω_i: the weighting coefficient for each loss component.

position-aware head. Unlike a standard YOLO head, this design contains four parallel prediction branches for (a) bounding box regression, (b) class prediction, (c) objectness score, and (d) our novel h_{rel} prediction. This additional branch outputs a normalized h_{rel} [0, 1] value for each detected tomato, supervised by the h_{rel} attribute defined during annotation, which precisely quantifies the vertical position of each fruit within its cluster.

(3) Rank-aware joint supervision: The model is trained end-to-end using a composite loss function. While the bounding box, class, and objectness branches utilize standard YOLO loss, the h_{rel} and class prediction are jointly supervised by the proposed DM-RankLoss. This loss explicitly encodes the biological ripening sequence by constructing pairwise ranking constraints, enforcing that fruits with higher ripeness levels occupy higher physical positions on the cluster, thereby integrating domain knowledge directly into the optimization objective.

2.2.2 Efficient position-aware head

Conventional object detection heads, such as those employed in early YOLO versions, typically predict both object class and location within a shared convolutional layer. This coupled design often leads to optimization conflicts between different tasks. Modern high-performance detectors (e.g., YOLOv8) have proven that adopting a "decoupled head" paradigm, where classification and regression are handled by separate branches, improves performance by mitigating spatial misalignment and enabling task-specific optimization. However, even these decoupled heads remain incapable of capturing relative topological relationships among objects.

To address this limitation while adhering to the well-established decoupled design principle, we propose an efficient position-aware head. Its core innovation lies in introducing a fourth parallel branch dedicated to predicting h_{rel} integrated into a mature decoupled-head architecture. The detailed structure is shown in Fig. 5.

Engineered for maximum computational efficiency without sacrificing performance, this module is well-suited for deployment in resource-constrained edge environments. Its design aligns closely with cutting-edge rapid target detection technologies transforming agriculture (Chu et al., 2025), and involves two key techniques:

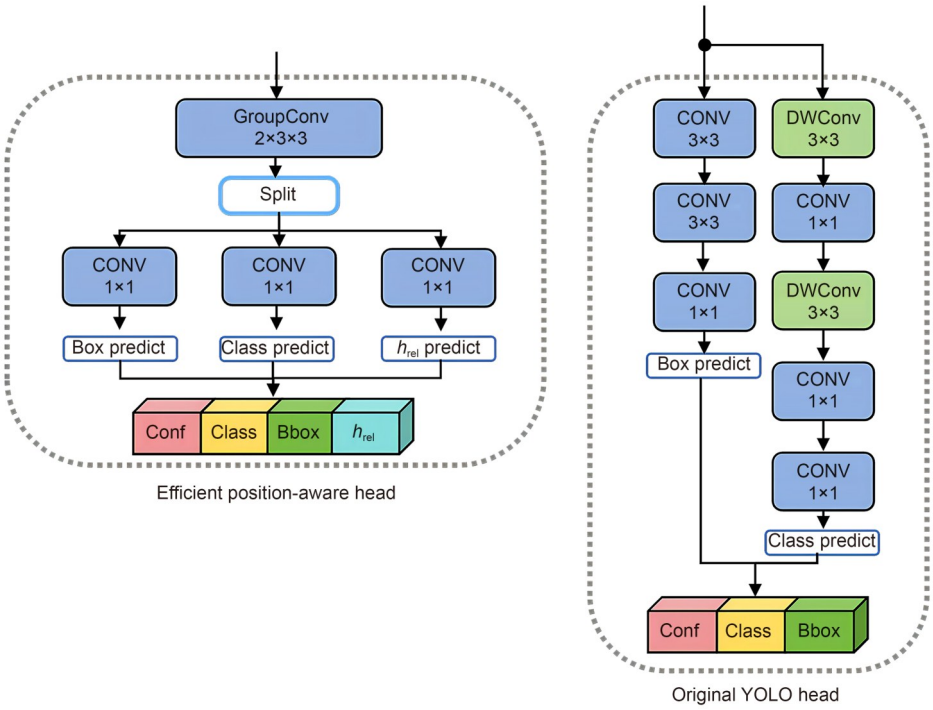

Fig. 5 Structural comparison between the proposed efficient position-aware head and a standard YOLO detection head. The enhanced design incorporates a fourth parallel branch for h_{rel} prediction and utilizes lightweight group convolutions. GroupConv: grouped convolution; CONV: convolutional layer; DWConv: depthwise convolution; Conf: confidence score representing the probability of object existence and localization accuracy; Bbox: bounding box; h_{rel}: normalized relative height within the cluster.

(1) Lightweight feature extraction via group convolutions (Krizhevsky et al., 2017): The feature map from the neck network initially passes through a shared stem consisting of two 3×3 group convolution layers. Unlike standard convolution (Fig. 5), group convolution splits input channels into independent groups, conducting convolution within each group separately. This approach substantially reduces both parameter count and computational complexity (floating point operations (FLOPs)), while preserving representation capabilities, thereby serving as a fundamental lightweighting technique.

(2) Fully decoupled parallel prediction branches: The refined feature map is then split and directed into three fully independent prediction branches. Each branch consists of a single 1×1 convolution layer that is responsible for distinct prediction tasks.

Through this design, all prediction tasks are structurally decoupled (including objectness, class, bounding box, and h_{rel}). Compared to the original multi-layer convolutional structure head (Fig. 5), our architecture is not only more interpretable but also computationally efficient, which is well-suited for edge computing.

2.2.3 Design of loss functions

Tomatoes in the same cluster exhibit a distinct ripening pattern due to physiological and environmental factors: fruits located at higher positions mature earlier than those lower in the cluster. The h_{rel} of a fruit serves as a critical topological attribute, which is essential for assessing cluster ripeness and predicting yield as crucial aspects of agricultural productivity. Conventional detection losses, such as cross-entropy and bounding box regression losses, fail to explicitly capture such structural relationships among objects. To incorporate this prior knowledge, we propose a hybrid loss function named DM-RankLoss that is specifically designed for fruit ranking tasks. Unlike approaches that rely on individual height values, DM-RankLoss dynamically adjusts the ranking criteria based on the relative ordering of all fruits within a cluster. Inspired by MarginRankingLoss, which is commonly used in machine learning for learning relative orders, our loss incorporates a dynamic margin tailored to the characteristics of fruit ripening sequences, enhancing the ability of the model to prioritize salient features for accurate ordinal prediction.

For a cluster C containing n_C detected fruits, the objective is to predict a relative height (h_i^{pred}) for each fruit i, such that the ordinal relations among predictions align with the ground-truth relative heights (h_i^{gt}). The loss builds upon the pairwise probability modeling framework from Learning-to-Rank (Burges et al., 2005) and emphasizes metric consistency in a manner similar to LambdaRank/LambdaMART (Burges, 2010). We further introduce physically informed pair construction and a dynamic margin term to enhance the discrimination between adjacent ranks and challenging samples, thereby adapting the loss to the spatial-ordinal structure of fruit clusters.

The DM-RankLoss consists of three key components: a base pairwise ranking loss L_{rank}, a differentiable re-weighting penalty loss L_{penalty}, and an auxiliary regression loss L_{reg}. The total loss is defined as the mean loss over all valid fruit pairs within a batch ($L_{\text{DM-rank}}$):

$$L_{\text{DM-rank}} = \frac{1}{N_{\text{p}}} \sum_{C \in \text{Batch}} \Big(\sum_{(i,j) \in P_C} \big(L_{\text{rank}}(i,j) + L_{\text{penalty}}(i,j) \big) + \lambda L_{\text{reg}}(C) \Big), \tag{3}$$

where P_C denotes the set of all unique fruit pairs (i, j) in cluster C, N_{p} represents the total number of fruit pairs in the batch, and λ is a hyperparameter that balances the contribution of the regression loss.

Conventional ranking losses commonly employ a fixed margin M. However, we posit that when two fruits are widely separated on a cluster, the difference in their predicted h_{rel} should be more pronounced. A fixed margin fails to capture this nuanced relationship; therefore, we propose a dynamic margin mechanism. For any fruit pair (i, j) within a cluster, the dynamic margin M_{ij} is defined proportionally to the difference in their ground-truth h_{rel}:

$$M_{ij} = M_{\text{base}} + \alpha \cdot \left| h_i^{\text{gt}} - h_j^{\text{gt}} \right|, \tag{4}$$

where M_{base} denotes a base margin ensuring discriminability even for adjacent fruits, and α is a scaling factor modulating the influence of height differences. L_{rank} is thus formulated as:

$$L_{\text{rank}}(i,j) = \max\left(0, -y_{ij} \cdot (h_j^{\text{pred}} - h_i^{\text{pred}}) + M_{ij}\right), \tag{5}$$

where $y_{ij}=\text{sign}(h_j^{\text{gt}}-h_i^{\text{gt}})$ indicates the true ranking direction. If $h_j^{\text{gt}}>h_i^{\text{gt}}$, then $y_{ij}=1$, and the loss encourages $h_j^{\text{pred}}>h_i^{\text{pred}}$, and vice versa. This dynamic margin design not only enforces correct ordinal relationships but also promotes a discriminative embedding space reflective

of true spatial distances. This idea draws inspiration from margin-based techniques in feature learning. For example, ArcFace used a fixed angular margin to enhance inter-class separation (Deng et al., 2019), while CurricularFace adaptively weights easy and hard samples during training (Huang et al., 2020).

A persistent challenge in tomato ripeness detection is distinguishing between fully ripe and breaking-stage fruits due to their color similarity, as highlighted in recent studies (Wang et al., 2024). To enhance the sensitivity of the model to spatial relationships between different maturity stages, we developed a differentiable, class-aware re-weighting penalty module. A simple conditional penalty would render the loss non-differentiable and impede gradient backpropagation. Instead, we leverage the smooth property of the Sigmoid function to construct a differentiable "degree of error" score.

Let us consider a critical error case: $h_i^{\mathrm{gt}}<h_j^{\mathrm{gt}}$, where fruit i is fully ripe and fruit j is ripe, but the model predicts $h_i^{\mathrm{pred}}\geqslant h_j^{\mathrm{pred}}$ (i.e., i is incorrectly predicted to be below j). The penalty term L_{penalty} is then defined as:

$$L_{\mathrm{penalty}}(i,j)=\beta\cdot\sigma\left(k\cdot\left(h_i^{\mathrm{pred}}-h_j^{\mathrm{pred}}\right)\right)\cdot\mathbb{I}\left(y_{ij}=1,\ \mathrm{cls}_i=0,\mathrm{cls}_j=1\right),\tag{6}$$

where $\sigma(x)$ denotes the Sigmoid function $\sigma(x)=1/(1+e^{-x})$, k is a hyperparameter controlling the steepness of the Sigmoid function, β is a positive scaling factor that modulates the maximum penalty strength, and $\mathbb{I}$ represents the indicator function, which outputs 1 when the specified conditions (y_{ij}=1, cls_i=0, cls_j=1) are simultaneously met and 0 otherwise. As the prediction error increases (i.e., ($h_i^{\mathrm{pred}}-h_j^{\mathrm{pred}}$) grows), $\sigma(\cdot)$ approaches 1 and the penalty tends toward β. If the predicted order is correct, the penalty term vanishes. This creates a smooth, adaptive penalty that scales with the severity of the ranking error and the class combination, guiding the model to prioritize correcting the most critical ranking errors.

Loss function based solely on ranking may preserve ordinal relationships and allow predicted h^{pred} values to drift globally. To anchor predictions within [0, 1] and provide more stable gradients, we introduce an auxiliary Smooth L1 regression loss (L_{reg}):

$$L_{\mathrm{reg}}(C)=\frac{1}{n_C}\sum\nolimits_{i=1}^{n_C}\mathrm{SmoothL1}\left(h_i^{\mathrm{pred}},h_i^{\mathrm{gt}}\right).\tag{7}$$

This regression term complements the ranking loss, ensuring an accurate estimation of both absolute and h_{rel}. Together, these three components form the DM-RankLoss, which effectively supervises the network to learn fine-grained relative spatial relationships between objects, significantly enhancing the perception of topological structure in complex scenes.

2.3 Model training

Detailed training configurations are provided in Method S3.

The model was evaluated on the basis of the following metrics: precision (P), recall (R), average precision (AP), mAP, F1-score (F_1), height-mean absolute error (H-MAE), and average inference time.

Here, precision is defined as the proportion of true positives (TP) among all samples, reflecting the accuracy of positive predictions. Recall, calculated as positive samples to the total number of actual positive samples (TP+false negatives (FN)), measures the ability of the model to identify all positive instances. AP, the area under the precision–recall curve, indicates detection accuracy, with values near 1 denoting excellent performance. The mAP computes the mean AP across all classes. The F1-score balances precision and recall as their harmonic mean. H-MAE quantifies the average absolute deviation between the predicted and ground-truth h_{rel} of the fruits. The corresponding formulas are listed below:

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\tag{8}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\tag{9}$$

$$\mathrm{AP}=\int_0^1 P(R)\,\mathrm{d}R,\tag{10}$$

$$\mathrm{mAP}=\frac{1}{n}\sum\nolimits_{i=1}^{n}\mathrm{AP}_i,\tag{11}$$

$$F_1=\frac{2PR}{P+R},\tag{12}$$

$$\text{H-MAE}=\frac{1}{N}\sum\nolimits_{i=1}^{N}\left|h_i^{\mathrm{pred}}-h_i^{\mathrm{gt}}\right|,\tag{13}$$

where N is the number and FP is false positive.

Furthermore, to directly evaluate the ability of the model to learn the spatial ordinal relationships of the fruits, we introduce a new metric: pairwise ranking accuracy (PRA), which measures how well the predicted heights preserve the true top-to-bottom ranking within a fruit cluster. For a cluster of "n" fruits, we first

sort them by their ground-truth relative height (h^{gt}). We then assess whether the predicted heights, h^{pred}, follow the same non-decreasing order. The PRA is calculated as follows:

$$\mathrm{PRA}=\frac{1}{N}\sum_{C=1}^{N}\left[\frac{1}{n_C-1}\sum_{j=1}^{n_C-1}\mathbb{I}\left(h_{j+1}^{\mathrm{pred}}\geqslant h_j^{\mathrm{pred}}\right)\right], \quad (14)$$

where N represents the total number of valid clusters in the validation set. The term h_j^{pred} represents the predicted height of the jth fruit in the sorted sequence. A higher PRA value (closer to 100%) reflects a more precise and accurate understanding of the topological structure within fruit clusters.

To assess the robustness and reproducibility of the proposed improvements, we conducted statistical significance testing. The baseline YOLOv12 and the proposed rank-aware YOLOv12 were trained five times, each with different random seeds to account for stochastic variations arising from weight initialization. To assess the statistical significance of the observed performance improvements, a paired t-test was subsequently applied to the resulting mAP at intersection over union (IoU) threshold of 0.50 (mAP_{50}) values.

3 Results

3.1 Generalization ability and applicability

To evaluate the generalizability of the proposed approach, the rank-aware module was integrated into four mainstream detection models: YOLOv8-n, YOLOv9-t, YOLOv11-n, and YOLOv12-n. As summarized in Table 1, all augmented models exhibited consistent improvements. Specifically, the mAP_{50} increased by 4.96, 4.59, 5.49, and 5.66 percentage points for the respective models, while each achieved a PRA exceeding 80%. Moreover, the mAP_{50} values of the rank-aware YOLO models were significantly higher than those of their baseline counterparts (P<0.05), indicating that the observed improvements were statistically significant and not attributable to random variations in training. In addition to accuracy gains, the integration also improved computational efficiency, reducing parameters by at least 10% and lowering inference latency. The enhanced models maintained a low H-MAE, consistently ranging between 0.10 and 0.14. As shown in Fig. 6, the improved model also exhibited faster and more stable loss convergence compared to the baseline YOLOv12.

3.2 Ablation study

To evaluate the individual contributions of each component within the rank-aware module, ablation studies were performed on the YOLOv12-n baseline.

3.2.1 Efficacy of the efficient position-aware head

We first introduced the h_{rel} prediction branch supervised by a standard Smooth L1 loss. As shown in Table 2 (Experiment 1 (Exp. 1) vs. Exp. 2), adding this branch improved the mAP_{50} from 92.53% to 94.21% (+1.68 percentage points), yielding an H-MAE of 0.203.

3.2.2 Efficacy of the DM-rank loss function

We then progressively replaced the Smooth L1 loss with the proposed DM-RankLoss components. Employing the full DM-RankLoss (Exp. 5) further boosted

Table 1 Performance comparison of different YOLO architectures before and after integration with the rank-aware module

Model	Parameters (M)	GFLOPs	mAP_{50} (%)	$mAP_{50:95}$ (%)	H-MAE	Latency (ms)	PRA (%)
YOLOv8-n	3.15	8.7	90.03	67.66		2.83	
YOLOv8-n (improved)	2.45	5.7	94.99*	68.32*	0.114	2.02	83.28
YOLOv9-t	2.09	8.2	89.98	65.42		2.55	
YOLOv9-t (improved)	**1.53**	5.4	94.57*	67.97*	0.132	1.81	80.10
YOLOv11-n	2.61	6.5	91.28	65.45		1.86	
YOLOv11-n (improved)	2.34	**5.2**	96.77*	66.20*	0.118	**1.65**	82.61
YOLOv12-n	2.57	6.5	92.53	68.20		1.89	
YOLOv12-n (improved)	2.30	5.3	**98.19***	**69.98***	**0.107**	1.66	**84.59**

* Significant differences between the improved YOLO models and their corresponding baselines, P<0.05. Bold values indicate the best performance. Parameters (M): number of parameters (millions); GFLOPs: giga floating point operations; mAP_{50}: mean average precision at intersection over union (IoU) threshold of 0.50; $mAP_{50:95}$: mean average precision at IoU thresholds from 0.50 to 0.95; H-MAE: height-mean absolute error; Latency (ms): inference time in milliseconds; PRA: pairwise ranking accuracy.

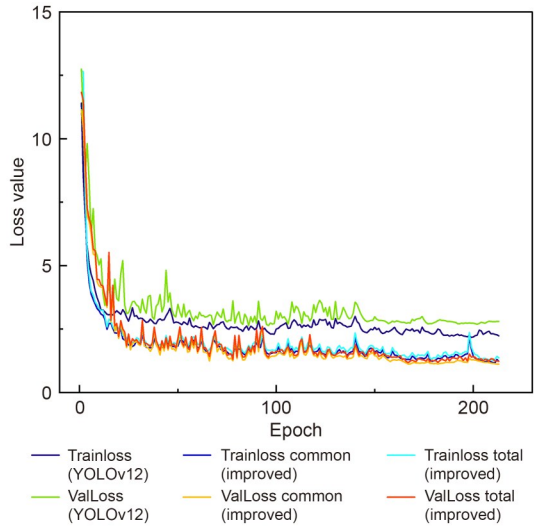

Fig. 6 Training and validation loss curves for the baseline YOLOv12 and the improved model. The loss for the improved model is separated into "common loss" (shared with baseline) and "total loss" (including dynamic margin-aware ranking loss (DM-RankLoss)).

the mAP_{50} to 98.19% and significantly reduced the H-MAE to 0.107. The corresponding PRA value also rose from 60.09% to 84.59%. As shown in Fig. 7, the DM-RankLoss and PRA exhibited a strong negative correlation during training. In the first 20 epochs, the PRA value surged from approximately 50% to over 70% as the loss decreased rapidly, eventually stabilizing around 85%.

3.3 Qualitative analysis

Visualizations were generated to provide an intuitive depiction of the model's learning behavior in challenging scenarios.

3.3.1 Visualization of detection results

To visually compare the rank-aware module with the baseline model, we selected challenging test-set scenarios under simulated dark, low, and strong lighting conditions. Fig. 8 presents the detection results from the baseline and our improved model. In the dark-light scenario where fruit surface color information is partially missing, the baseline misclassified a breaking fruit on the far right as ripe and misclassified a ripe fruit as a fully ripe one. In contrast, using top fully ripe fruits as a spatial reference, our model correctly identified both. Under low light, where ripe and fully ripe fruits appeared similar in color, the baseline misclassified several bottom-ripe fruits as fully ripe, while our model accurately recognized their ripeness stages. In strong light, the baseline confused a breaking fruit with a green fruit, leading to a misjudgment of the entire cluster as mature, an error that our model successfully avoided.

To quantitatively validate the robustness suggested by the visual analysis, we evaluated the models on test subsets categorized by illumination conditions (Table 3). The rank-aware model demonstrated consistent improvements across all lighting scenarios. Most notably, under the challenging "dark" condition where visual texture features are severely degraded, our method achieved a pronounced improvement of 8.33% in the mAP_{50}. This quantitative result strongly supports our hypothesis that the encoded topological information functions as a critical compensatory feature when RGB-based appearance cues become unreliable.

3.3.2 Visualization of h_{rel} feature learning

To verify whether the model can genuinely learn the concept of "h_{rel}," we visualized the predicted h_{rel} values directly on the detection results (Fig. 9), following explainable artificial intelligence (XAI) principles in object detection (Vondrick et al., 2016). The bounding box colors represent the h_{rel} values on a gradient scale (blue for high, yellow for low). A smooth color transition from top to bottom within each cluster

Table 2 Ablation study on the components of the rank-aware module

Experiment	h_{rel} branch	L_{rank}	$L_{penalty}$	L_{reg}	mAP_{50} (%)	ΔmAP_{50} (%)	H-MAE	PRA (%)
1	×	×	×	×	92.53			
2	√	×	×	×	94.21	+1.68	0.203	60.09
3	√	√	×	×	96.39	+2.18	0.188	71.19
4	√	√	√	×	97.11	+0.72	0.162	83.60
5 (ours)	√	√	√	√	98.19	+1.08	0.107	84.59

h_{rel} branch: vertical relative height branch; L_{rank}: pairwise ranking loss; $L_{penalty}$: differentiable re-weighting penalty loss; L_{reg}: auxiliary regression loss; mAP_{50}: mean average precision at intersection over union (IoU) threshold of 0.50; H-MAE: height-mean absolute error; PRA: pairwise ranking accuracy. √ indicates that the component is used; × indicates that the component is not used.

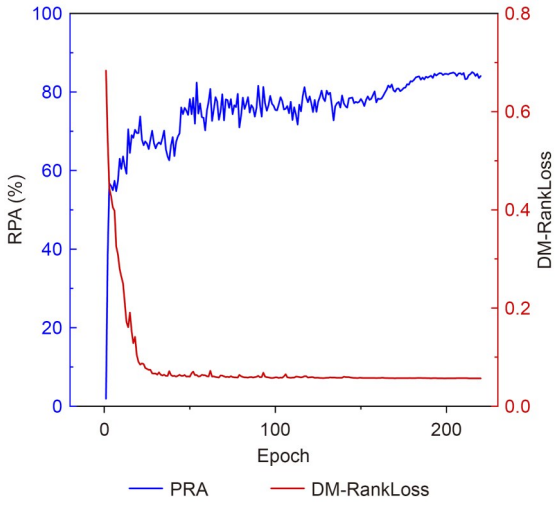

Fig. 7 Relationship between dynamic margin-aware ranking loss (DM-RankLoss) and pairwise ranking accuracy (PRA) on the validation set during training. A strong negative correlation is observed between the two metrics.

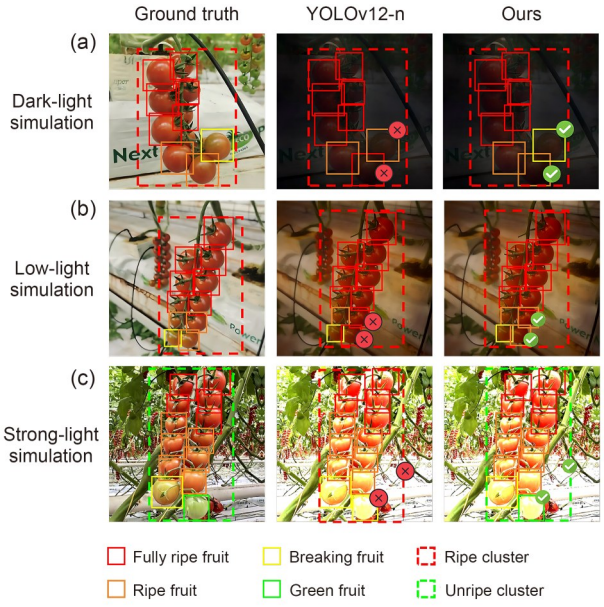

Fig. 8 Qualitative comparison of detection robustness under challenging lighting conditions. (a), (b), and (c) correspond to simulated dark, low, and strong lighting environments, respectively. The baseline YOLOv12 produces several classification errors (marked with red crosses): In (a), it misclassifies a "breaking" fruit as "ripe" and a "ripe" fruit as "fully ripe"; In (b), it labels two dimly lit "ripe" fruits as "fully ripe"; In (c), strong specular reflections cause it to miss an "unripe" fruit, resulting in the erroneous classification of the entire unripe cluster as "ripe." In contrast, the proposed rank-aware model (ours) corrects all these errors. Notably, in (b), it correctly identifies the bottom fruit as "ripe" despite an ambiguous color by leveraging the spatial prior that distal fruits are typically less mature than proximal ones.

Table 3 Performance comparison under different lighting conditions

Test subset	mAP_{50}(%)		
	YOLOv12-n	Rank-aware YOLOv12 (ours)	Improvement (Δ)
Dark light	81.42	89.75	+8.33
Low light	84.60	91.20	+6.60
Strong light	83.15	88.90	+5.75

mAP_{50}: mean average precision at intersection over union (IoU) threshold of 0.50.

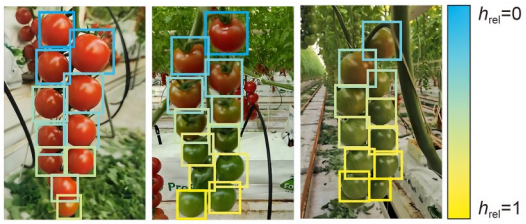

Fig. 9 Visualization of the learned vertical relative height (h_{rel}) features. Bounding boxes are color-coded according to the regression output of the position-aware head, with a continuous gradient from blue (top, h_{rel}=0) to yellow (bottom, h_{rel}=1). The smooth and consistent color transition across clusters demonstrates that the model has effectively learned the physical structure of tomato spikes, independent of visual ripeness features.

is observed, consistent with the actual spatial arrangement of fruits.

3.3.3 Confusion matrix

To gain a deeper insight into the performance improvement brought by the rank-aware module, we compared the normalized confusion matrices of the baseline (YOLOv12-n) and our final model (Fig. 10). The baseline model (Fig. 10a) exhibited a 14% misclassification rate between the visually similar "fully ripe" and "ripe" classes. Our rank-aware model reduced this confusion to a negligible level (Fig. 10b), demonstrating a clear advantage in classification accuracy.

4 Discussion

This study introduces and validates a novel approach for integrating domain (the spatial ripening sequence of fruits) into a deep learning-based object detection framework. The experimental results demonstrate that the proposed method not only outperforms baseline models in detecting the ripeness of on-cluster

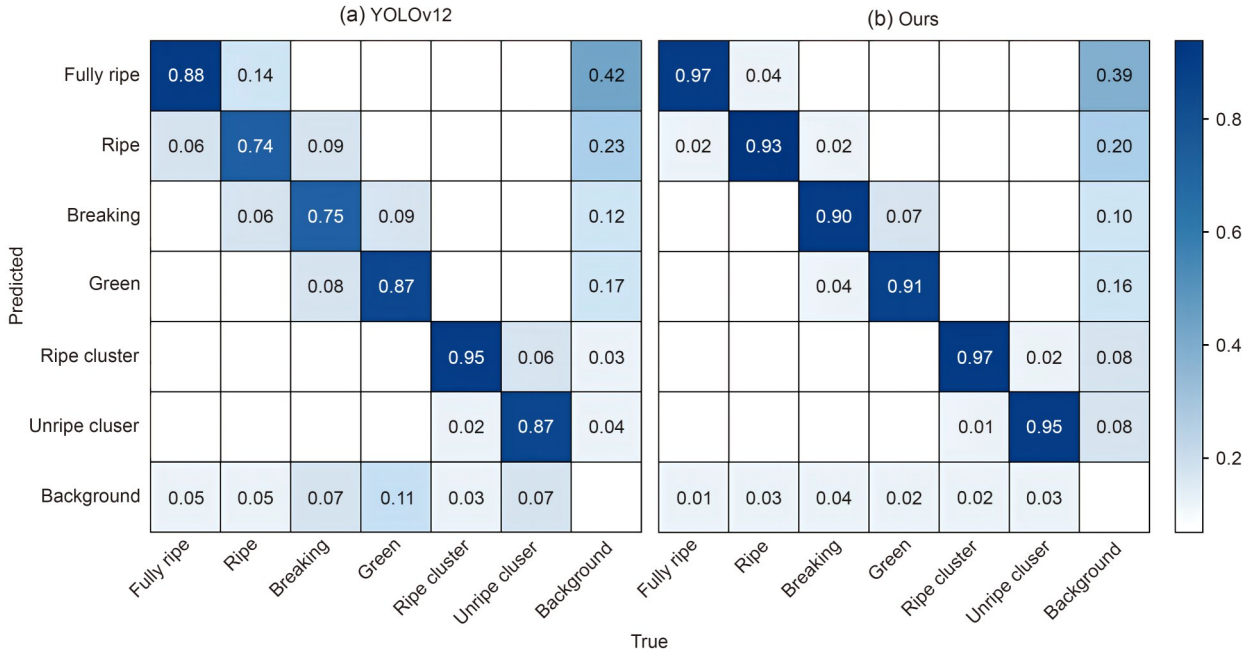

Fig. 10 Normalized confusion matrix comparing the classification performance of the baseline YOLOv12 model (a) and the proposed rank-aware YOLO (b). Values represent normalized classification accuracy (0–1). Diagonal entries indicate correct predictions; off-diagonal entries reflect confusion between ripeness stages. (a) The baseline model shows moderate diagonal dominance, with correct rates ranging from 0.74 to 0.95. It exhibits notable confusion between adjacent ripeness stages: green vs. breaking (0.17), breaking vs. ripe (0.15), and ripe vs. fully ripe (0.20). (b) The rank-aware model achieves stronger diagonal dominance, with correct classification rates between 0.90 and 0.97. Adjacent-stage confusion is markedly reduced: green–breaking confusion decreases from 0.17 to 0.11, breaking–ripe confusion decreases from 0.15 to 0.02, and ripe–fully ripe confusion decreases from 0.20 to 0.06. These results validate that incorporating spatial rank information effectively mitigates color ambiguity in ripeness classification.

tomatoes but also offers a viable strategy for improving model robustness in complex environments. This section discusses the main findings, their implications, and the limitations of this work, along with potential future directions.

4.1 Interpretation of findings: mechanism of the rank-aware module

Our principal finding indicates that explicitly modeling the spatial–ordinal relationship among tomatoes within a cluster effectively overcomes the limitations inherent to purely appearance-based recognition methods. The effectiveness of the rank-aware module stems from the synergistic effect of its architectural design and its dedicated learning objective (Table 1). Although adopting the YOLOv12 architecture improves overall efficiency through its residual efficient layer aggregation network (R-ELAN) backbone and FlashAttention mechanisms (Dao et al., 2022), a substantial and distinct portion of parameter reduction can be directly attributable to our proposed efficient position-aware head. While standard YOLO detection heads typically rely on stacked 3×3 convolutions to decouple features, our design utilizes lightweight group convolutions and streamlined 1×1 regression branches. This targeted structural optimization reduces the parameter count of the detection head by approximately 40% compared to the conventional YOLOv12 head.

The ablation studies we performed provided clear evidence for this mechanism. Simply adding an h_{rel} prediction branch yielded a notable 1.72% mAP_{50} improvement. This indicates that even basic spatial awareness helps the model distinguish visually similar objects. The most significant gain, however, came from the introduction of DM-RankLoss. When fully applied, it nearly halved the H-MAE (from 0.203 to 0.107), confirming our core hypothesis: for this task, learning the correct ordinal relationship is more critical and effective than regressing an exact metric value. The ranking-based supervision guides the model to learn a more robust representation of the cluster's spatial layout, which is particularly crucial in dense arrangements where preserving

the top-to-bottom sequence matters more than slight coordinate inaccuracies. Furthermore, the training dynamics revealed a regularizing effect of DM-RankLoss. The consistent reduction in the "common loss" of our improved model (Fig. 6) suggests that the structured, physics-informed constraint of the ripening sequence helps the network learn more discriminative features, which in turn facilitates the optimization of the primary detection tasks (classification and localization), demonstrating that ranking and detection are mutually beneficial.

The practical impact of this mechanism is clearly reflected in the qualitative results. The confusion rate between the "fully ripe" and "ripe" classes dropped from 14% to a negligible level (Fig. 10), a direct consequence of the model's acquired spatial prior. In challenging lighting conditions where local color cues become unreliable (Fig. 8), our model leverages its learned topological knowledge, reasoning that a lower fruit is statistically less likely to be fully ripe than a similar-looking fruit above it. This ability of our model to resolve ambiguity using contextual, structural information—rather than relying solely on ambiguous local cues—forms the foundation of its improved robustness observed in our experiments.

4.2 Comparison with prior work and broader implications

Previous studies have sought to improve robustness by incorporating additional hardware such as depth or hyperspectral cameras (Chen WJ et al., 2024). For instance, recent studies have utilized hyperspectral imaging for non-destructive tomato quality assessment (Fass et al., 2025), as well as advanced ground robots integrating multi-spectral, thermal, and depth cameras for high-throughput phenotyping (Su et al., 2025). Although effective, such hardware-intensive approaches face scalability barriers. In contrast, this study introduces a novel, software-centric paradigm. Our approach shows that embedding domain-specific priors directly into the learning process can yield substantial performance improvements using only standard RGB images. This strategy has profound practical implications, offering a cost-effective pathway toward highly robust and deployable vision systems for agricultural robotics, eliminating the need for expensive and often fragile multi-modal sensors.

Theoretically, this work aligns with an emerging principle in computer vision and artificial intelligence (AI): integrating structural or physical constraints can shift models from pure pattern recognition to structured reasoning (Banerjee et al., 2020; Wang et al., 2025). By explicitly encoding domain knowledge, the model can learn not only visual features but also spatial and ordinal relationships. For example, Banerjee et al. (2020) incorporated topological priors from Discrete Morse theory to improve neuron connectivity in segmentation tasks. Similarly, Kumar et al. (2024) applied topological regularization to maintain structural consistency in generated light detection and ranging (LiDAR) point clouds, while Wang et al. (2025) integrated shadow formation physics to enhance shadow removal. Our work extends this paradigm of research by encoding a well-established biological principle as a soft constraint for robust agricultural phenotyping. Moreover, our network design leverages densely connected components, a strategy also effective in other complex domains such as glioma segmentation in medical imaging (Zhang et al., 2021), underscoring the general applicability of our architectural choices.

4.3 Limitations and failure analysis

Despite the remarkable performance improvements, it is important to delineate the operational boundaries of the proposed framework. As illustrated in Fig. 11, two primary failure modes have been identified. First, the topological prior of the method depends on accurately inferring the global structure of the tomato cluster. Under conditions of severe occlusion (Fig. 11a), the model loses its spatial reference, leading to a systematic bias in the predicted h_{rel} and consequent ripeness misclassification. Second, while the model addresses color ambiguity caused by lighting, detection in densely overlapping clusters remains a challenge (Fig. 11b). When clusters are tightly packed, the model often fails to reliably delineate instance boundaries, merging multiple clusters into one. This ambiguity compromises h_{rel} estimation and results in incorrect maturity predictions.

Beyond these model-specific failures, the study has two main experimental limitations. First, the dataset was collected from a single tomato variety under specific greenhouse environments. Thus, generalization to other cultivars with differing morphology and maturation patterns remains to be applied. Second, our approach approximates three-dimensional (3D) spatial relationships using 2D vertical coordinates from a single viewpoint. Although effective for typical vertical clusters, this simplification may fail under severe

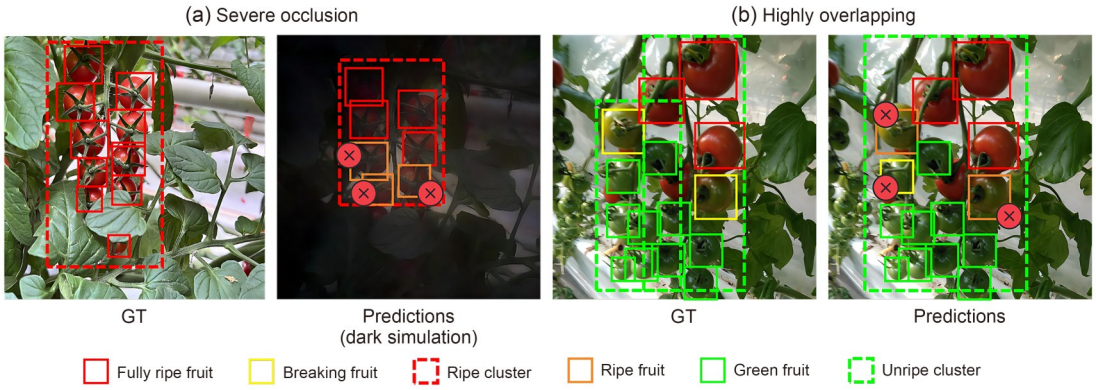

Fig. 11 Typical failure cases illustrating the robustness boundaries of the rank-aware model. (a) Severe occlusion: heavy foliage occlusion at the bottom end disrupts the topological anchor, leading to erroneous classifications of the fruits (red cross). (b) High overlap: in highly clustered scenarios, severe overlap between adjacent clusters causes bounding box regression ambiguity, leading to incorrect maturity prediction (red cross). GT: ground truth.

occlusion, cluster distortion, or atypical viewpoints, where 2D projection loses critical 3D information. In future work, we will extend the rank-aware framework to operate directly on 3D point clouds. This direction aligns with advances in 3D plant phenotyping using RGB-D (red, green, blue, and depth) data for structural reconstruction (Sampaio et al., 2021) and represents a natural progression, given that our data were acquired with a stereo depth camera.

5 Conclusions

To address the limitations of current object detection models in overlooking physical topological relationships, particularly in fine-grained agricultural recognition tasks under visual ambiguity, this study proposes and validates a novel rank-aware YOLO framework. Its core is a plug-and-play "rank-aware module," comprising an efficient position-aware head and a DM-RankLoss, which explicitly encodes the spatial prior that on-cluster tomatoes ripen from top to bottom into the model training process. Experiments on tomato ripeness detection confirmed the effectiveness of our approach:

(1) Performance and generalizability: After integrating the rank-aware module, several state-of-the-art detectors (YOLOv8-n, YOLOv9-t, YOLOv11-n, and YOLOv12-n) showed significant accuracy improvements, with mAP_{50} increases of 4.96, 4.59, 5.49, and 5.66 percentage points, respectively. This demonstrates the universality of our method as a model-agnostic enhancement.

(2) Efficacy of core components: Ablation studies highlighted the importance of DM-RankLoss. Compared to a standard regression loss, this ranking-based loss more effectively guided the model to learn object spatial layout, raising mAP_{50} from 94.21% to 98.19%, and nearly halving the H-MAE (from 0.203 to 0.107).

(3) Underlying mechanism: Visualizations confirmed that the model learns to accurately predict h_{rel} and uses this spatial information to resolve ambiguous ripeness stages that are difficult to distinguish by color alone.

The implications of this work extend beyond tomato detection. Practically, it offers a cost-effective path toward robust vision systems for harvesting robots, without relying on expensive multi-modal sensors. Theoretically, it provides a simple yet effective paradigm for embedding structural priors to improve computer vision models.

Future work will focus on: (1) extending the framework to other crops with similar spatial-ordinal patterns (e.g., grapes, peppers) and exploring more complex topological relationships; and (2) advancing from 2D to 3D spatial modeling using previously collected OAK-D stereo data, which should help resolve pose-related ambiguities by working in 3D point cloud space. We believe that integrating real-world physical knowledge is a vital step toward advancing AI from perception to cognition.

Data availability statement

Data will be made available upon request from the authors.

Acknowledgments

This work was supported by the Science and Technology Program of the Ministry of Agriculture and Rural Affairs of the People's Republic of China.

Author contributions

Bingquan CHU and Ruiyuan WU wrote the original manuscript. Fengle ZHU and Yong HE conceived the study, with Yong HE also providing supervision. Bingquan CHU, Ruiyuan WU, and Haijun ZHANG performed the investigation and software development. Formal analysis was conducted by Haijun ZHANG and Haochuan QIN, and validation was performed by Ruiyuan WU, Haochuan QIN, and Zishun PENG. Yong HE and Fengle ZHU reviewed and edited the manuscript. Bingquan CHU was responsible for data and funding acquisition, while Yong HE handled data curation and provided resources with Fengle ZHU. All authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Bingquan CHU, Ruiyuan WU, Haijun ZHANG, Haochuan QIN, Zishun PENG, Fengle ZHU, and Yong HE declare that they have no conflicts of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

Declaration on the use of generative AI tools

During the preparation of this manuscript, the authors used DeepSeek solely for the purpose of grammar and spelling checks. The authors reviewed and edited all content and take full responsibility for the final publication.

References

Albahar M, 2023. A survey on deep learning and its impact on agriculture: challenges and opportunities. *Agriculture*, 13(3):540. https://doi.org/10.3390/agriculture13030540

All China Federation of Supply and Marketing Cooperatives, 2021. Tomato, GH/T 1193-2021. All China Federation of Supply and Marketing Cooperatives, China.

Banerjee S, Magee L, Wang DK, et al., 2020. Semantic segmentation of microscopic neuroanatomical data by combining topological priors with encoder–decoder deep networks. *Nat Mach Intell*, 2(10):585-594. https://doi.org/10.1038/s42256-020-0227-9

Burges C, Shaked T, Renshaw E, et al., 2005. Learning to rank using gradient descent. Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, p.89-96. https://doi.org/10.1145/1102351.1102363

Burges CJC, 2010. From RankNet to LambdaRank to LambdaMART: an overview. Microsoft Research Technical Report, MSR-TR-2010-82. Available from: https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview

Cao WZ, Mirjalili V, Raschka S, 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recogn Lett*, 140:325-331. https://doi.org/10.1016/j.patrec.2020.11.008

Chen WB, Liu MC, Zhao CJ, et al., 2024. MTD-YOLO: multi-task deep convolutional neural network for cherry tomato fruit bunch maturity detection. *Comput Electron Agric*, 216: 108533. https://doi.org/10.1016/j.compag.2023.108533

Chen WJ, Rao Y, Wang FY, et al., 2024. MLP-based multimodal tomato detection in complex scenarios: insights from task-specific analysis of feature fusion architectures. *Comput Electron Agric*, 221:108951. https://doi.org/10.1016/j.compag.2024.108951

Chu BQ, Guo ZY, Liu BJ, et al., 2025. Fast detection of rice striped stem borer (*Chilo suppressalis*) stress based on UAV sensor and multimodal segmentation method. *Plant Growth Regul*, 105(4):1057-1071. https://doi.org/10.1007/s10725-025-01320-8

Dao T, Fu DY, Ermon S, et al., 2022. FlashAttention: fast and memory-efficient exact attention with IO-awareness. arXiv: 2205.14135. https://doi.org/10.48550/arXiv.2205.14135

Deng JK, Guo J, Xue NN, et al., 2019. ArcFace: additive angular margin loss for deep face recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, p.4685-4694. https://doi.org/10.1109/CVPR.2019.00482

Díaz R, Marathe A, 2019. Soft labels for ordinal regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, p.4733-4742. https://doi.org/10.1109/CVPR.2019.00487

Fass E, Shlomi E, Ziv C, et al., 2025. Machine learning models based on hyperspectral imaging for pre-harvest tomato fruit quality monitoring. *Comput Electron Agric*, 229:109788. https://doi.org/10.1016/j.compag.2024.109788

Food and Agriculture Organization of the United Nations (FAO), 2023. Crops and livestock products. https://www.fao.org/faostat/en/#data/QCL [Accessed on Oct. 1, 2025].

Fu H, Gong MM, Wang CH, et al., 2018. Deep ordinal regression network for monocular depth estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, p.2002-2011. https://doi.org/10.1109/CVPR.2018.00214

Gautier H, Rocci A, Buret M, et al., 2005. Fruit load or fruit position alters response to temperature and subsequently cherry tomato quality. *J Sci Food Agric*, 85(6):1009-1016. https://doi.org/10.1002/jsfa.2060

Huang YG, Wang YH, Tai Y, et al., 2020. CurricularFace: adaptive curriculum learning loss for deep face recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, p.5900-5909. https://doi.org/10.1109/CVPR42600.2020.00594

Khan Z, Shen Y, Liu H, 2025. ObjectDetection in agriculture: a comprehensive review of methods, applications, challenges, and future directions. *Agriculture*, 15(13):1351.

https://doi.org/10.3390/agriculture15131351

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. https://doi.org/10.1145/3065386

Kumar P, Bhat KM, Shenvi Nadkarni VB, et al., 2024. GLiDR: topologically regularized graph generative network for sparse LiDAR point clouds. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, p.15152-15161. https://doi.org/10.1109/CVPR52733.2024.01435

Li RZ, Ji ZJ, Hu SK, et al., 2023. Tomato maturity recognition model based on improved YOLOv5 in greenhouse. *Agronomy*, 13(2):603. https://doi.org/10.3390/agronomy13020603

Li XX, Chen WB, Wang YQ, et al., 2023. Design and experiment of an automatic cherry tomato harvesting system based on cascade vision detection. *Trans Chin Soc Agric Eng*, 39(1): 136-145 (in Chinese). https://doi.org/10.11975/j.issn.1002-6819.202210099

Sampaio GS, Silva LA, Marengoni M, 2021. 3D reconstruction of non-rigid plants and sensor data fusion for agriculture phenotyping. *Sensors*, 21(12):4115. https://doi.org/10.3390/s21124115

Su M, Zhou D, Yun YZ, et al., 2025. Design and implementation of a high-throughput field phenotyping robot for acquiring multisensor data in wheat. *Plant Phenomics*, 7(2):100014. https://doi.org/10.1016/j.plaphe.2025.100014

Vondrick C, Khosla A, Pirsiavash H, et al., 2016. Visualizing object detection features. *Int J Comput Vision*, 119(2):145-158. https://doi.org/10.1007/s11263-016-0884-7

Wang AC, Qian WH, Li A, et al., 2024. NVW-YOLOv8s: an improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages. *Comput Electron Agric*, 219:108833. https://doi.org/10.1016/j.compag.2024.108833

Wang XR, Guo LQ, Wang XY, et al., 2025. SoftShadow: leveraging soft masks for penumbra-aware shadow removal. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA, p.23217-23226. https://doi.org/10.1109/CVPR52734.2025.02162

Wang Z, Ling YM, Wang XL, et al., 2022. An improved Faster R-CNN model for multi-object tomato maturity detection in complex scenarios. *Ecol Inform*, 72:101886. https://doi.org/10.1016/j.ecoinf.2022.101886

Xiao F, Wang HB, Xu YQ, et al., 2023. Fruit detection and recognition based on deep learning for automatic harvesting: an overview and review. *Agronomy*, 13(6):1625. https://doi.org/10.3390/agronomy13061625

Yao J, Ke XB, Gu XY et al., 2025. Optimized substrate selection for enhanced orchid growth based on high-throughput lysimetric arrays. *J Zhejiang Univ-Sci B*, online first. https://doi.org/10.1631/jzus.B2500195

Zhang XB, Hu Y, Chen W, et al., 2021. 3D brain glioma segmentation in MRI through integrating multiple densely connected 2D convolutional neural networks. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 22(6):462-475. https://doi.org/10.1631/jzus.B2000381

Supplementary information

Methods S1–S3