Computers and Electronics in Agriculture 252 (2026) 112095

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

LSCDNet: leveraging fine-grained contextual cues for precise detection of small target crop disease spots in complex field environments

Mengyao Ma [a], Yanfen Li [a,*], Jipei Cao [a], Hanxiang Wang [a], Tan N. Nguyen [b], L. Minh Dang [c,d,*]

[a] School of Computer Science, Qufu Normal University, Rizhao 276826, China
[b] Department of Architectural Engineering, Sejong University, Seoul 05006, the Republic of Korea
[c] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam
[d] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Viet Nam

ARTICLE INFO

Keywords:
Crop disease detection
Small target object localization
Context-aware feature fusion
Multi-scale feature representation

ABSTRACT

Timely crop disease monitoring is essential for stable grain production and quality improvement, with accurate identification of small, visually similar disease spots still a key challenge in complex field environments. The traditional disease detection methods are inefficient, high in cost, and highly dependent on experts' experience for support. This study introduces an advanced YOLO framework for precise crop disease detection, which consists of three core improved modules: Firstly, a novel convolution structure, SSRConv, is proposed to adaptively capture the irregular textures and topological changes of crop disease spots with extremely low computational overhead, thereby achieving high-robustness feature extraction in resource-constrained agricultural scenarios. Secondly, the proposed C2F-S2Block module adopts a dual-branch structure with star convolution as its core. It enhances cross-scale feature dependencies, enabling precise characterization of crop diseases while effectively suppressing background interference. Thirdly, we design the CLDA-Head detection module, which integrates cross-layer attention and distribution-aware regression to enhance feature fusion accuracy and bounding box localization stability. Further leveraging the enhanced CLLABlock and SimAM attention mechanisms, the module simultaneously improves multi-scale perception and classification capabilities. Experimental results on the large-scale crop disease dataset DCrop12 show that the proposed LSCDNet model achieves 52.6 % mean average precision (mAP_{50}) at a detection rate of 77 FPS, representing an improvement of 2.7 percentage points over the baseline model. The proposed model outperforms existing state-of-the-art methods and establishes a new benchmark for real-time crop disease detection in complex field scenarios.

1. Introduction

The progress of human society has always been rooted in agriculture, and the stable development of agriculture is an important prerequisite for ensuring the orderly advancement of social economy. However, the current process of agricultural development faces with numerous restrictive factors, among which the problems in the field of crop disease detection are particularly prominent. Relevant research data show that various major crops worldwide are generally affected by compound pests and diseases, leading to a significant decline in yield, causing substantial and unavoidable losses to agricultural production, and seriously affecting the stability and sustainability of agricultural production (Savary et al., 2019; Zhang, 2023). Faced with these challenges, the importance of accurate detection of crop diseases is evident to ensure the steady development of agricultural harvest.

In the field of crop disease monitoring, traditional detection methods rely on expert visual identification, which can hardly meet the requirements of large-scale modern agricultural production (Jogekar and Tiwari, 2020; Dang, 2024). With the continuous rapid advancement of machine learning (ML) technologies and the ongoing development of smart agriculture, crop object detection has ushered in a crucial era of technological innovation. Among diverse machine learning algorithms, support vector machines (SVM) have garnered widespread attention for their outstanding classification performance, whereas decision tree algorithms have been extensively applied in disease identification tasks within smart agricultural systems (Lai et al., 2024). Although traditional

* Corresponding authors.
E-mail addresses: yanfen@qfnu.edu.cn (Y. Li), danglienminh@duytan.edu.vn (L.M. Dang).

https://doi.org/10.1016/j.compag.2026.112095
Received 30 August 2025; Received in revised form 22 April 2026; Accepted 20 June 2026
Available online 26 June 2026

machine learning models can perform automated feature extraction and classification using annotated datasets, thereby improving the accuracy and efficiency of crop detection to a certain extent, they still exhibit obvious limitations. Such models generally suffer from insufficient detection precision, perform poorly in complex and dynamic agricultural environments, and are characterized by cumbersome feature extraction processes as well as weak adaptability to crop diversity and dynamics.

To address the aforementioned limitations, classic convolutional neural networks (CNNs) typified by ResNet and DenseNet are capable of automatically extracting hierarchical features from crop images through deep network architectures. On this basis, the disease recognition model based on multi-branch fine-grained learning, with ResNet-50 as the backbone, achieves improved accuracy in tomato disease identification and grading through object positioning, destruction and reassembly, as well as attention area division (Zhang, 2023). For apple disease detection, the multi-model fusion and voting decision strategy based on MobileNet, Xception and InceptionResNet effectively improves the detection accuracy and generalization ability for multi-class and multi-label diseases on apple leaves (Vora and Padalia, 2210). Nevertheless, despite the remarkable progress achieved by CNNs in detection efficiency and feature representation, as crop disease detection increasingly demands higher accuracy and adaptability to complex field environments, CNNs show limitations in capturing long-range dependencies and holistic features. Against this backdrop, the Vision Transformer (ViT) model, based on the Transformer architecture, has emerged as a promising solution for crop image analysis (Azad et al., 2024). By leveraging self-attention mechanisms, ViT effectively processes global information and autonomously learns hierarchical features from large-scale crop images. However, ViT models still suffer from high computational complexity and insufficient accuracy in capturing fine-grained local features. To address these shortcomings, recent state-of-the-art (SOTA) models like Real-Time Detection Transformer (RT-DETR) combine Transformer's global feature capability with real-time detection performance (Zhao, 2024). As a representative Transformer-based detector, it optimizes the encoder-decoder structure and adopts dynamic query generation, achieving a better balance between accuracy and speed than traditional DETR models. However, for crop disease detection, which involves small scattered disease spots, variable field lighting, and deployment on low-power agricultural terminals, RT-DETR still lacks sufficient sensitivity to fine-grained disease features and has relatively high computational overhead for on-site use. Given these challenges at different stages of detection technology development, it is crucial to develop a novel algorithm that balances detection accuracy and efficiency.

Focusing on the core demand for lightweight crop disease detection, this study selects YOLOv12 as the baseline model. To improve crop disease detection performance, key contributions are proposed from two dimensions: network structure optimization and dataset support.

The core findings and contributions of this study are as follows:

(1) A novel convolution structure termed SSRConv is proposed to adaptively capture irregular textures and topological variations of crop disease spots with extremely low computational overhead. It breaks the bottleneck between model size and detection accuracy in existing lightweight networks, enabling highly robust feature extraction in resource-constrained agricultural scenarios.
(2) The proposed C2F-S2Block module centers on a dual-branch architecture and star operation. It compensates for the shortcomings of traditional feature fusion methods, enhances cross-scale correlation, captures subtle crop disease features, and effectively suppresses background interference.
(3) The designed CLDA-Head integrates cross-layer attention with distribution-aware regression, thereby improving feature fusion accuracy and bounding box localization stability. Additionally, the enhanced CLLABlock and SimAM attention mechanisms further strengthen multi-scale perception and classification capabilities.
(4) This study constructed a large-scale dataset containing 136,443 images, covering 6 dicotyledon crops susceptible to fungal diseases and 4 typical high-impact fungal diseases. Each crop corresponds to 2 diseases, forming a total of 12 disease categories, which provides authentic and reliable data support for related research on crop disease detection.

The organizational structure of the subsequent chapters is as follows. Section 2 reviews the relevant studies conducted in this specific field. Section 3 focuses on the constructed model and the employed methods. Section 4 elaborates on the experimental process implemented for the given problem and presents the corresponding results. Finally, Section 5 summarizes the research findings, including a comparative analysis of different models, and discusses potential areas for future exploration in this research direction.

2. Related work

2.1. Lightweight object detection networks

Manual inspection is time-consuming and error-prone (Li, 2020), while CNN-based R-CNN methods suffer from high computational costs (Magdy, 2025), making lightweight object detection networks a research hotspot in agriculture. Representative lightweight models such as YOLO (Redmon et al., 2016), Tiny-YOLO (Redmon and Farhadi, 1804), and MobileNet have been widely applied, but MobileNet series still perform suboptimally in small-scale object detection (Howard et al., 1704). To address this, Tianping Li et al. proposed the Spatial Group-wise Enhance (SGE) module, which enhances semantic feature expression and suppresses noise by generating attention factors for each spatial location in each semantic group (Li et al., 1905).

Although existing lightweight object detection networks have attained promising performance in crop disease detection tasks, they still struggle with inherent limitations in detection accuracy, generalization capacity across complex field scenarios, and robustness to low-quality or imbalanced field data. To alleviate these issues, this study introduces a novel lightweight convolution module termed SSRConv, as elaborated in Section 3.1. In contrast to conventional lightweight convolution structures, SSRConv organically integrates spatial rearrangement, depthwise separable convolution (DSC), channel shuffle, and a lightweight residual compensation branch. By explicitly strengthening the extraction of fine-grained and weak disease features that are easily neglected in real scenes, SSRConv effectively compensates for the inadequate capability of traditional modules in capturing subtle disease spot patterns and low-contrast visual characteristics. Meanwhile, the proposed architecture maintains high computational efficiency while achieving superior feature representation, thereby realizing a more favorable trade-off between model lightweightness, detection speed, and identification precision.

2.2. Multi-scale feature fusion modules

In crop disease detection, significant scale variation of disease spots represents a core challenge, rendering multi-scale feature fusion a key technique for improving detection performance (Shafay, 2025). Although existing studies have introduced early image pyramid-based approaches and various CNN-based fusion modules (Lin, 2017), including FPN (Ross and Dollár, 2017); BiFPN, Efficient RepGFPN (Xu et al., 2211), LFPN (Xie et al., 2022), the improved PANet by Roy et al. (Roy and Bhaduri, 2022; Roy et al., 2022), and attention mechanism-based methods (Xue and Marculescu, 2023; Wang et al., 2020), current FPN-based variants inevitably increase algorithmic complexity. Conventional fusion strategies fail to effectively model complex nonlinear relationships among cross-scale features and are

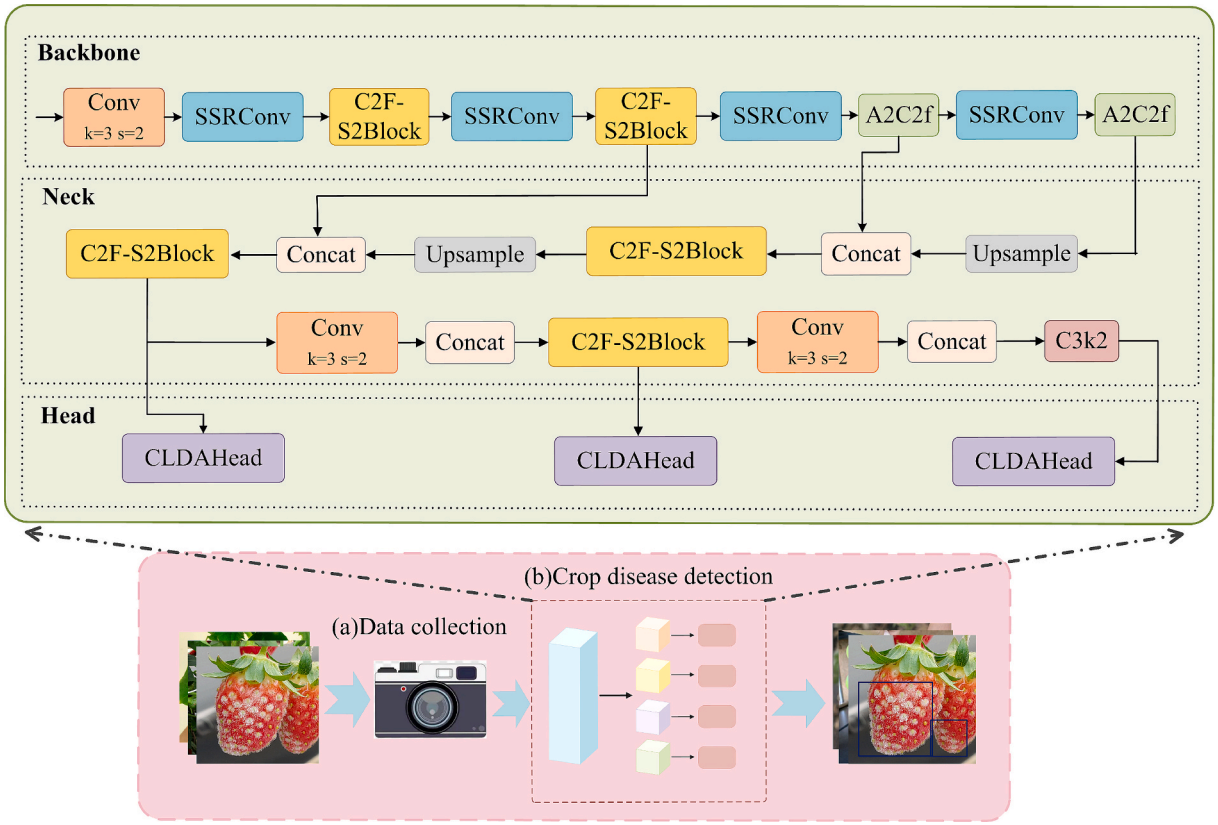

Fig. 1. Depiction of the advanced hybrid framework for crop disease detection (LSCDNet).

restricted by fixed-weight mechanisms, which limits model performance in detecting small-scale disease spots.

Compared with conventional multi-scale feature fusion strategies, the proposed C2F-S2Block module, as elaborated in Section 3.2, adopts a dual-branch structure centered on star operation, nonlinear feature interaction, and selective attention mechanism. By dynamically modeling complex nonlinear dependencies across different scales, this module effectively strengthens semantic correlations between multi-level features and significantly enriches the diversity and discrimination of disease representations. Moreover, it achieves a favorable trade-off between lightweight architectural design and powerful feature expression ability. By adaptively focusing on multi-scale and fine-grained crop disease features, the C2F-S2Block module suppresses irrelevant background noise and redundant information. Thus, the C2F-S2Block module effectively overcomes the limitations of traditional fusion methods in nonlinear relation modeling, scale adaptation, and the capture of weak and subtle disease spot characteristics.

2.3. Micro-sample detection heads

As a key task in computer vision and machine learning, micro-sample detection faces significant challenges in various domains (Li et al., 2024), and researchers have proposed multiple solutions for small-object detection in crop disease detection. For example, Cheng et al. (Dai et al., 2021), Roy et al. (Roy and Bhaduri, 2023), and other studies (Jianqiang, 2024; Zhou et al., 2024) have developed improved detection frameworks or heads to enhance small-object detection performance, but these methods still have limitations in adaptability or computational efficiency.

However, in practical crop disease detection scenarios, existing detection heads still face the key challenge of balancing computational efficiency and small object detection accuracy. To address these challenges, as elaborated in Section 3.3, this study introduces the proposed CLDA–Head, which effectively improves the accuracy of feature fusion and the stability of bounding box localization, enabling it to exhibit excellent robustness and detection performance in small object detection scenarios.

3. Methodology

In this study, the overall workflow of the proposed crop disease detection framework is illustrated in Fig. 1. Based on image data, the framework comprises three core modules: a feature extraction backbone, a feature fusion module, and a detection head. Specifically, SSRConv (marked in blue) is deployed in the backbone network, C2F-S2Block (marked in yellow) in the neck network, and CLDA-Head (marked in purple) in the detection head. All modules are clearly distinguished via visual coding. In the training stage, input images are initially processed by the feature extraction module to obtain rich semantic representations. The detailed design of the network architecture will be elaborated in Section 3.1. Subsequently, the features extracted from the initial feature extraction stage are fed into the feature pyramid for further processing. For a detailed description of the specific architecture and operational mechanism of the feature pyramid, please refer to Section 3.2. Ultimately, the head structure designed in this study undertakes the core tasks of object recognition and localization within the images. Further details are provided in Section 3.3. Upon completion of training, the model's weight parameters are persistently saved as files for subsequent testing.

3.1. Lightweight and efficient convolutional module

The SSRConv module aims to reduce computational cost while preserving fine-grained features and enhancing the model's sensitivity to subtle visual patterns. It achieves this by integrating spatial rearrangement, DSC, the channel shuffle mechanism, and a lightweight residual

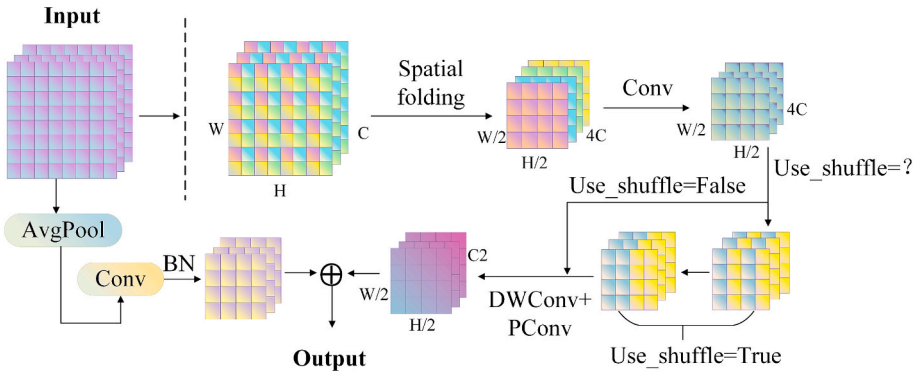

Fig. 2. Illustration of the proposed SSRConv module, which integrates spatial compression and channel reconstruction to enhance receptive field and efficiency.

compensation branch.

The overall structure of the SSRConv module is illustrated in Fig. 2. First, the Spatial folding operation rearranges spatial information into the channel dimension, achieving an information-preserving downsampling strategy. This operation effectively retains fine-grained structures in the image and offers a significant advantage in representing features of small object regions. In the implementation, for an input feature map with dimensions $H \times W \times C$, after applying Spatial folding with a downsampling factor r, the resulting feature map is reshaped to dimensions of $\frac{H}{r} \times \frac{W}{r} \times (C \cdot r^2)$. The computation can be formulated by Equation (1):

$$X_{folding} = \text{Pixel-Folding}_r(X_{in}) \tag{1}$$

where $X_{folding} \in \mathbb{R}^{B \times C \cdot r^2 \times \frac{H}{r} \times \frac{W}{r}}$. This architecture reduces the spatial resolution while retaining a greater amount of original fine-grained detail, which facilitates the extraction of richer contextual features by subsequent network modules. Given a shuffle group number g, the channel dimension of the input feature is initially sectioned into g groups, and then the channels within each group are rearranged according to a predefined rule. Through this shuffling operation, the feature interaction and information fusion between different channel groups are effectively enhanced.

To effectively reduce the computational complexity of the model and enhance the efficiency of feature modeling, this work adopts DSC in place of conventional convolution operations. This structure decouples spatial feature extraction and channel fusion, thereby significantly decreasing the number of parameters and computational overhead of the network. To mitigate the potential semantic information loss caused by spatial rearrangement and separable convolutions, a lightweight residual compensation path is designed to provide additional feature support. This branch first performs spatial downsampling on the input feature X using a two-dimensional average pooling operation. Subsequently, a 1×1 convolution is applied to project the features to the target channel dimension, followed by batch normalization (BN) for normalization. The final output is the compensated feature as shown in Equation (2):

$$Y_{skip} = BN(W_{skip} * X_{pool}), W_{skip} \in \mathbb{R}^{C' \times C \times 1 \times 1} \tag{2}$$

This compensation branch not only captures global contextual information but also maintains a lightweight structure, thereby improving the overall semantic integrity and representation capability of the module. The final output is generated by element-wise addition of the main branch and the residual compensation branch, followed by a SiLU nonlinear activation. Based on the above design, the SSRConv module effectively realizes substantial feature compression and computational acceleration.

3.2. A feature interaction and discrimination-oriented C2F block

The proposed improved model offers distinct advantages in feature representation and fusion efficiency: it enhances cross-scale feature

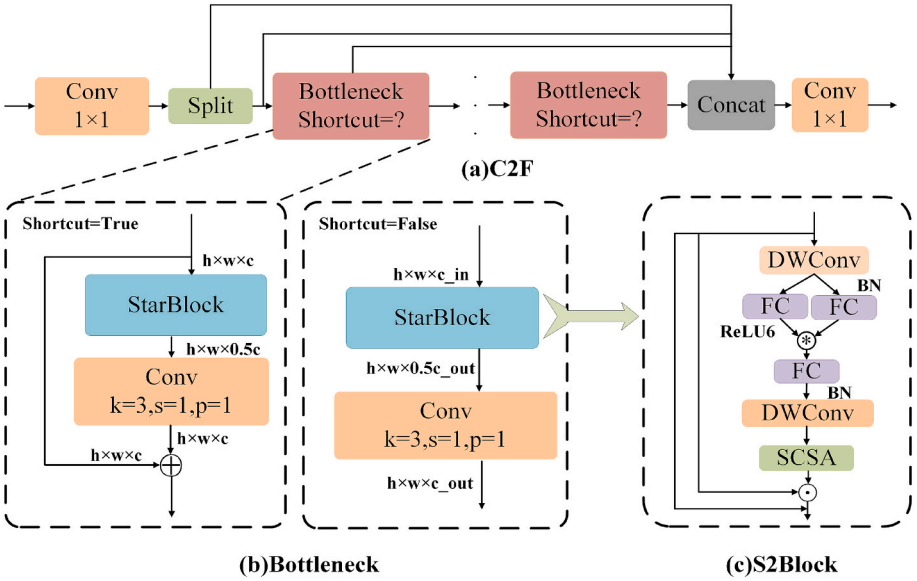

Fig. 3. The architecture of C2F-S2Block and the proposed S2Block. FC refers to Fully Connected layer, and SCSA stands for Spatial-Channel Self-Attention.

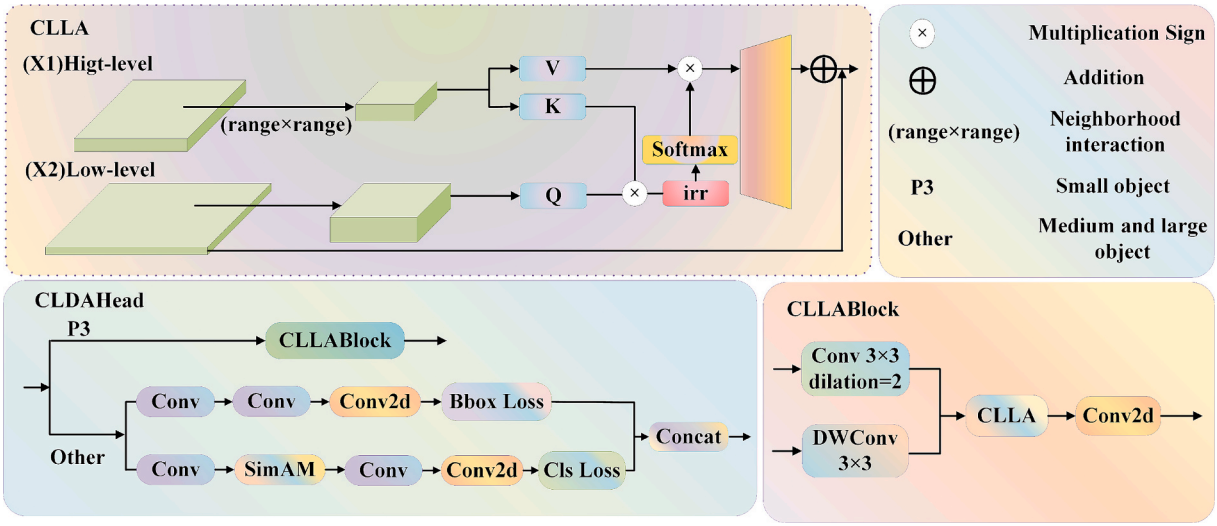

Fig. 4. The overall architecture presentation of our CLDA-Head. CLLA is a channel-wise attention mechanism with local context awareness.

correlation, enables hierarchical semantic integration, and enriches feature diversity via a nonlinear interaction mechanism—supplying ample discriminative cues for object detection in complex scenarios. We propose integrating the StarBlock into the C2F block (depicted in Fig. 3(a)). The C2F-StarBlock module replaced all instances of the C3K2 module in YOLOv12, with the goal of enhancing feature representation diversity and improving the modeling of nonlinear interactions. The proposed improvement strategy is mainly inspired by findings from research on StarNet (Ma, 2024). Specifically, we substitute the first convolutional layer in the Bottleneck unit of the C2F block with a StarBlock to enhance feature extraction capabilities. The StarBlock achieves feature fusion via a dual-branch convolutional structure combined with an element-wise multiplication operation (referred to as the star operation), which allows the input features to be implicitly projected into a high-dimensional nonlinear space. This enhances the network's capacity to model complex patterns while maintaining a low parameter overhead.

The StarBlock structure adopts a dual-branch design. Firstly, the input features pass through a DW-Conv layer followed by BN to extract local spatial information. Subsequently, the features are independently processed by two fully connected layers, generating the feature representations for the dual branches, denoted as $W_1^T X$ and $W_2^T X$ respectively. The nonlinear mapping is introduced in one branch by applying the ReLU6 activation function. The outputs of the two branches are then fused through element-wise multiplication, referred to as the star operation, and formulated as Equation (3):

$$\mathrm{Star}(X) = \mathrm{act}(W_1^T X) \ * \ (W_2^T X) \tag{3}$$

where $*$ indicates element-wise multiplication. This operation allows for the projection of input features into a high-dimensional nonlinear space without expanding the channel dimension, thereby enhancing the expressiveness of the features while preserving architectural compactness. The fused features are further processed by a linear layer to align the output dimensionality, and then refined through a DW-Conv layer, which facilitates feature reconstruction and the extraction of fine-grained details. In the StarBlock, the star operation is typically formulated as Equation (4):

$$(W_1^T X + B_1) \ * \ (W_2^T X + B_2) \tag{4}$$

which represents the element-wise multiplication-based fusion of two linearly transformed feature representations. We concatenate the transformation weights and bias parameters into a single entity, denoted as $W = \begin{bmatrix} W \\ B \end{bmatrix}$, and similarly represent the input as $X = \begin{bmatrix} X \\ 1 \end{bmatrix}$ Under this formulation, the star operation can be simplified to $(W_1^T X) * (W_2^T X)$. As an illustrative example, consider a scenario with a single-element input and one output channel. Let $w_1, w_2, X \in \mathbb{R}^{(d+1)\times 1}$, where d represents the input channel count. This can be extended to multiple output channels, where W_1, W_2, $X \in \mathbb{R}^{(d+1)\times(d'+1)}$, and multiple feature elements are processed simultaneously with $X \in \mathbb{R}^{(d+1)\times n}$.

The star operation can be reformulated in Equation (5) to (9):

$$w_1^T x * w_2^T x \tag{5}$$

$$= \left(\sum_{i=1}^{d+1} w_1^i x^i\right) * \left(\sum_{j=1}^{d+1} w_2^j x^j\right) \tag{6}$$

$$= \sum_{i=1}^{d+1}\sum_{j=1}^{d+1} w_1^i w_2^j x^i x^j \tag{7}$$

$$\underbrace{= \alpha_{(1,1)} x^1 x^1 + \cdots + \alpha_{(4,5)} x^4 x^5 + \cdots + \alpha_{(d+1,d+1)} x^{d+1} x^{d+1}}_{\frac{(d+2)(d+1)}{2} \text{ items}} \tag{8}$$

$$\alpha(a,b) = \begin{cases} w_1^a w_2^b, & \text{if } a = b, \\ w_1^a w_2^b + w_1^b w_2^a, & \text{if } a \neq b. \end{cases} \tag{9}$$

Here, a and b index the channels, while α denotes the coefficient assigned to each item. In this way, the star operation is expanded into $\frac{(d+2)(d+1)}{2}$ distinct terms. Except for $\alpha_{(d+1,:)} x^{d+1} x$, each term is a nonlinear function of x, implying that they correspond to independent implicit dimensions. In the above formulation, when $d \gg 2$, the reformulated operation maps the d-dimensional input into an implicit feature space of approximatel $\frac{(d+2)(d+1)}{2} \approx \left(\frac{d}{\sqrt{2}}\right)^2$ dimensions, thereby expanding the feature representation without introducing additional computational overhead.

Furthermore, to further improve the discriminative power and spatial sensitivity of the features, the Selective Channel Spatial Attention (SCSA) mechanism is incorporated after the second DW-Conv within the StarBlock module (Si, et al., 2024). The complete architecture of the StarBlock module integrated with this mechanism is illustrated in Fig. 3 (c), where the attention mask generated by SCSA is fused with the original input feature map via Hadamard product to enhance the feature representation of salient regions. Specifically, the SCSA module is

designed by integrating local convolution with multi-scale global convolution to capture spatial attention under different receptive fields, while employing a self-attention mechanism to generate channels. The spatial attention branch employs horizontal and vertical average pooling, combined with multi-scale DW-Conv to extract both local and global information, which are then fused via a gating mechanism such as the Sigmoid function. The channel attention is generated based on the downsampled feature maps using a channel-wise self-attention mechanism. The introduced SCSA fusion strategy enhances the model's feature representation capability while effectively alleviating the vanishing gradient problem during training.

3.3. Cross-layer distribution-aware attention head

Recent years have seen wide application of object detection in complex real-world environments such as autonomous driving, smart agriculture, and surveillance. However, detected objects vary greatly in scale, shape, and background interference, posing major challenges to existing frameworks. To tackle these issues, this paper proposes a novel detection head architecture named Cross-layer Distribution-aware Attention Head, CLDA-Head for short, and the relevant structure is shown in Fig. 4. The core innovation lies in the integration of the CLLA and DFL modules, which enhances the network's discriminative ability and regression stability in complicated agricultural scenes.

Specifically, the CLLA module resolves semantic inconsistency and spatial misalignment to facilitate comprehensive integration of high-level semantic features and low-level spatial details. The DFL module enhances object localization precision by reformulating bounding box regression as discrete probability distribution regression. Meanwhile, the CLLABlock enables cross-scale feature interaction and fusion, strengthening multi-scale contextual information association. Collectively, these designs endow CLDA-Head with more expressive target representations in complicated agricultural scenes, improving overall detection performance.

Algorithm 1: CLLA Attention Mechanism.

Input: Feature map $x_1 \in R^{B\times C\times W_1\times H_1}$, $x_2 \in R^{B\times C\times W_2\times H_2}$ Local range R, Channel dimension C.
Output: Fused feature map $out \in R^{B\times C\times W_2\times H_2}$.
pad ← [R/2] − 1
x_1_padded← ZeroPad2d (x_1, padding = (pad, pad, pad, pad))
x_2_padded← PermuteAndReshape (x_2, (B, W_2, H_2, C, 1))
Local_windows ← []
for $i \in [0, R-1]$, j ∈ [0, R − 1] :
  window ← x_1_padded […, i::2, j::2][…,:W_2,:H_2] # Stride = 2, match x_2 size
  Local_windows.append (Reshape (window, (B, 1, W_2, H_2, C)))
Local_windows ← Concatenate (Local_windows, dim = 1)
Q ← Linear (x_2_padded, C)
K ← Linear (Local_windows, C)
V ← Linear (Local_windows, C)
dots ← (Q ⊙ K) / R → Sum(dim = 4)
avg_dots ← Avg (dots, dim = 1) → Reshape (…, 1, W_2, H_2)
irr ← 2avg_dots − dots # Inverse range-aware score
att_weights ← Softmax (irr, dim = 1)
out_intermediate ← Sum (V ⊙ Reshape (att_weights, (B, R^2, W_2, H_2, 1)), dim = 1)
out ← PermuteAndReshape (out_intermediate, (B, C, W_2, H_2)) # Restore shape
out ← (out + x_2) / 2 # Residual average fusion
Return out

The detailed implementation process of the CLLA module is formally presented in Algorithm 1. The CLLA module takes two inputs: a low-level feature map $X_1 \in \mathbb{R}^{B\times C\times W_1\times H_1}$ and a high-level feature map $X_2 \in \mathbb{R}^{B\times C\times W_2\times H_2}$. The low-level feature map contains rich spatial information, while the high-level feature map carries high-level semantic information. Here, B denotes the batch size, C is the number of channels, and W, H represent the spatial width and height, respectively. In order to enrich the semantic representation of X_2, a spatially aligned local region of size 2 × 2 is extracted from X_1 for each spatial location in X_2, enabling the construction of position-specific local context to guide feature enhancement. Initially, X_1 is zero-padded, and a sliding window with kernel size 2 × 2 and stride 2 is used to extract local patches that are spatially aligned with each location in X_2, thereby ensuring effective scale correspondence between feature levels. Next, each position in X_2 is transformed into a query vector Q through a linear projection, whereas the extracted local patches from X_1 are mapped into key and value vectors K and V by separate linear layers, respectively. This operation can be formally defined as Equations (10):

$$Q = F_Q{\cdot}X_2, K = F_K{\cdot}X_{1,local}, V = F_V{\cdot}X_{2,local} \tag{10}$$

To highlight the most informative and non-redundant key regions, we incorporate an irregularity-aware attention mechanism (Equation (11)).

$$\alpha_{i,j} = 2{\cdot}mean\left(q_i{\cdot}k_{i,*}\right) - \left(q_i{\cdot}k_{i,j}\right) \tag{11}$$

This approach enhances the modeling of spatial distribution irregularities by amplifying attention weights on low-correlation regions while suppressing high-frequency redundant areas. Then, the attention scores are normalized via a Softmax function to obtain the attention weights A, based on which a weighted summation over the value vectors v is computed as shown in Equation (12):

$$y_i = \sum_{j=1}^{r^2} A_{i,j} \cdot v_{i,j} \tag{12}$$

Subsequently, the resulting output y is fused with the original high-level feature map X_2 via a residual connection, with the final enhanced feature representation is expressed as Equation (13):

$$\widehat{X_2} = \frac{1}{2}(y + X_2) \tag{13}$$

The introduced Distribution-aware Focal Loss (DFL) module, originally proposed in VarifocalNet (Zhang, 2021), differs from traditional direct regression approaches. DFL formulates the prediction of bounding box coordinates as a discrete probability distribution regression task. By employing a differentiable integral operation to transform the discrete distributions into continuous values, it effectively improves the accuracy and stability of bounding box prediction. In the implementation of DFL, the predicted output for each anchor is first reshaped to [B, 4, K, A], and then transposed to [B, K, 4, A] to facilitate the application of the softmax operation along the discrete dimension K for each coordinate component. Subsequently, a fixed-parameter 1 × 1 convolution is applied to the softmax output to perform a weighted summation, which is equivalent to executing an integral operation over the discrete distribution. The weights of this convolution are initialized as an integer vector ranging from 0 to K − 1, and are set to be non-trainable to ensure that the operation functions purely as an integral over the discrete distribution. The final output is a set of continuous bounding box coordinates with a shape of [B, 4, A].

Moreover, in the proposed CLLABlock, low-level feature X_1 has rich spatial details but limited receptive field, which restricts its contextual capture. To mitigate this, we apply dilated convolution to X_1 for enhanced contextual awareness. For the high-level counterpart X_2, while it is rich in semantics, it suffers from redundancy, weak spatial details, and high computational cost. Therefore, depthwise separable convolution is introduced to improve representation efficiency and reduce overhead. This hierarchical collaborative feature processing boosts feature complementarity, sharpens the attention mechanism's response to key regions, and ultimately improves overall detection performance.

In the CLDA-Head, to improve the expressiveness of classification features, we incorporate the SimAM attention mechanism into the classification branch. This module infers 3D attention weights for each position in the convolutional feature map without introducing additional parameters, thereby enhancing the model's ability to focus on target category information. This design enables the regression and classification branches to achieve independent and efficient

Table 1
Statistics of the total number of category images, total number of instances, average, median, and range of instance counts for 12 crop disease classes in the DCrop12 dataset.

Crop disease name	Code	Number of images	Number of instances	Average of instances	Median of instances	Range of instances
Strawberry-botrytis cinerea	a1	16,082	20,372	1.26	1	1–14
Strawberry-powder mildew	a2	11,048	22,242	2.01	1	1–25
Cucumber-botrytis cinerea	a3	10,521	112,651	10.70	8	1–100
Cucumber-powder mildew	a4	15,007	158,164	10.53	8	1–97
Tomato-powder mildew	a5	13,233	63,592	4.80	4	1–32
Tomato-botrytis cinerea	a6	10,317	41,759	4.04	4	1–24
Chili pepper-anthracnose	a7	12,191	20,092	1.64	1	1–19
Chili pepper-powder mildew	a8	10,967	38,039	3.46	1	1–61
Paprika-powder mildew	a9	25,083	286,081	11.40	8	1–100
Paprika-damping off	a10	1,508	1,508	1	1	1–2
Grape-anthracnose	a11	3,016	22,545	7.47	6	1–76
Grape-botrytis cinerea	a12	7,469	104,512	13.99	12	1–100
Total		136,442	891,557			

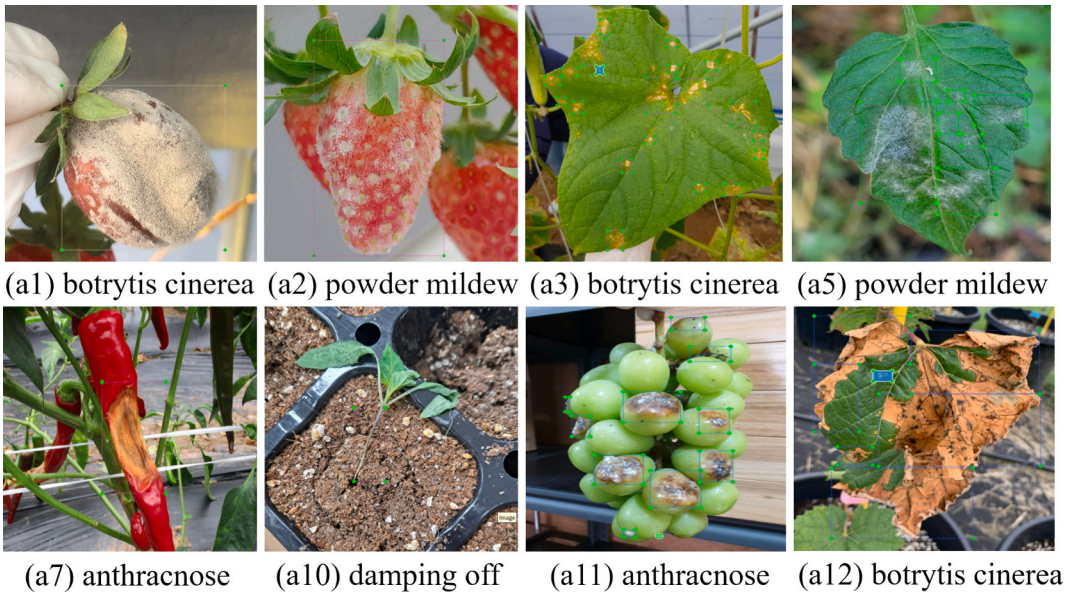

Fig. 5. 8 classes of the most representative crop diseases from the proposed DCrop12 dataset.

optimization for their respective tasks.

4. Experiment and result analysis

4.1. Dataset

1) Overview of the Dataset: In this work, the original data was provided by the National Information Society Agency of Korea (NIA). The data collection was a collaborative initiative led by Farm Hannong Company Limited, with active participation from multiple institutions, including Nonghyup University, Yonam College, Gyeongsangbuk-do Agricultural Research & Extension Services, and Jeonbuktechnopark. It should be noted that due to the constraints of the data collection context, potential biases may exist in the dataset: Such as geographical bias, and acquisition and annotation biases potentially arising from multi-institutional participation, which lead to inconsistencies in acquisition conditions and annotation standards. To address these biases, targeted optimizations have been implemented during the model design and training phases of this study. Based on the given large-scale dataset, an image compression technique was applied to reduce the image size. The compressed dataset is referred to as the Dicotyledon Crop Dataset (DCrop12), which contains a total of 136,443 images. From the perspective of crop species, the dataset covers six crops highly susceptible to fungal diseases, including cucumber, strawberry, grape, tomato, chili pepper, and paprika. Regarding disease types, it includes four typical fungal diseases: Anthracnose, Powdery Mildew, Botrytis cinerea, and Damping-off. These diseases essentially cover the major groups of common fungal diseases in dicotyledonous crops, exhibiting strong representativeness and providing solid support for the universality of the research conclusions. Furthermore, all selected diseases are those with the highest infection frequency, most severe damage, and greatest control difficulty under natural cultivation conditions, ensuring high compatibility between disease types and target crops. Each of the six common crops is associated with two specific disease types, covering a total of 12 disease types, labeled as a1 to a12. Table 1 presents a comprehensive overview of the DCrop12 dataset. Several representative samples from DCrop12 are shown in Fig. 5, which offers a deeper understanding of the nature and characteristics of the entire dataset. To ensure the smooth execution of the training, validation, and testing processes, the dataset was randomly partitioned at an 8:1:1 ratio. This random partitioning strategy effectively

Table 2
An ablation analysis was carried out involving the replacement of diverse components with the LSCDNet model.

Model	C2F-S2Block	CLDA-Head	SSRConv	mAP_{50}±(%)
YOLOv12n				49.9 ± 0.5
LSCDNet(1)	√			51.5 ± 0.4
LSCDNet(2)	√	√		51.8 ± 0.3
LSCDNet(3)	√	√	√	52.6 ± 0.3

ensures the independence of the training and testing sets, reduces potential data biases arising from fixed grouping schemes, and thereby further enhances the reliability of model evaluation results and the academic rigor of our work.

2) Data Preprocessing: Specifically, in its preprocessing stage, a proportional compression scheme was adopted to reduce the resolution of original 4 K images to the 1 K scale. The images are proportionally scaled to a target height of 1024 pixels: first, read the image width and height via the OpenCV library; then calculate the scaling factor (scaling factor = 1024/original height) and the corresponding target width (original width × scaling factor). Smooth scaling is finally achieved by combining one-step sampling (i.e., utilizing all original pixels without skipping any pixels) in both horizontal and vertical directions with bilinear interpolation. Despite the substantial reduction in image resolution, the compressed images still allow for satisfactory identification accuracy in crop disease identification, ensuring the feasibility and reliability of the study.

4.2. Training environment

In this study, all experiments were implemented based on the PyTorch 2.5.0 deep learning framework on the Ubuntu 20.04.1 operating system. The hardware platform consisted of an Intel(R) Xeon(R) Gold 5218N CPU (2.30 GHz), 256 GB of RAM, and four NVIDIA RTX A6000 GPUs (48 GB each), which supplied sufficient computational power. Uniform training hyperparameters were set for all comparison models. The input image resolution was fixed at 640 × 640 pixels. The initial learning rate was set to 0.01 using the SGD optimizer. The batch size was set to 32 to ensure efficient training, and the total number of training epochs was set to 100.

4.3. Ablation study

To investigate the impacts of various modules on the performance of the YOLOv12 model in plant disease recognition tasks, ablation experiments are designed and conducted in this section, with relevant results detailed in Table 2. Specifically, the core difference between LSCDNet (1) and the original YOLOv12 model lies in replacing the C3K2 module in the network with the C2F-S2Block module. This improvement increases the model's mean average precision (mAP_{50}) from the initial 49.9 % to 51.5 %, representing a 1.6 percentage point improvement. On this basis, LSCDNet (2) further upgrades the Detect head to the CLDA-Head, enabling the mAP_{50} metric to be further optimized to 51.8 %, achieving an additional 0.3 percentage point gain. Finally, for LSCDNet (3), the traditional Conv module is replaced with SSRConv, which effectively reduces the model's computational complexity and parameter overhead. Through the targeted improvements of the above series of network components, the mAP_{50} of the proposed LSCDNet model is 2.7 percentage points higher than that of the original YOLOv12 model, fully verifying the effectiveness and rationality of these improvement strategies in enhancing the model's plant disease detection performance.

4.4. Experiments on SSRConv

To evaluate the effectiveness of the proposed SSRConv, we conducted ablation experiments on the DCrop12 dataset. As shown in Table 3, considering the impact of FLOPs and parameter count on inference speed, we integrated SSRConv into the C2F block to reconstruct the backbone network. Specifically, the introduction of the Spatial folding module reduces FLOPs from 10.3 G to 9.6 G and parameters from 3.24 M to 3.09 M, demonstrating its direct effect in compressing computational complexity and parameter overhead. On this basis, the shuffle module further optimizes feature interaction, improving mAP_{50} from 51.7 % to 52.1 %, validating its enhancement of feature representation. Finally, the Residual module achieves a significant performance breakthrough, with Recall jumping to 49.1 % and mAP_{50} rising to 52.6 %. Its FLOPs increase slightly from 9.6 G to 9.7 G due to the additional identity mapping and shortcut connections in the residual structure, which introduce extra element-wise operations and feature fusion steps. However, by optimizing the gradient propagation path, this structure enables the model to learn hierarchical features more

Table 4
This table presents a comparative evaluation of various YOLOv12n-C2F variant models, including metrics such as Precision, Recall, mAP, parameter count, and computational cost (FLOPs).

Model	Precision (%)	Recall (%)	mAP_{50}± (%)	$mAP_{50\text{-}95}$ (%)	Param. (M)	FLOPs (G)
YOLOv12n-C2F-DSConv (Nascimento et al. 2019)	61.1	46.9	50.9 ± 0.6	28.6	3.42	11.0
YOLOv12n-C2F-MLCA (Wang, 2024)	60.7	47.1	50.7 ± 0.5	28.5	2.91	9.3
YOLOv12n-C2F-EMA (Ouyang, 2023)	60.6	47.0	50.8 ± 0.4	28.6	2.91	9.3
YOLOv12n-C2F-DBB (Ding, 2021)	60.8	47.6	51.1 ± 0.7	28.7	3.28	11.1
YOLOv12n-C2F-ScConv (Li et al., 2023)	59.2	46.2	49.4 ± 0.5	27.4	2.69	8.1
YOLOv12n-C2F + C3K2 (Owor, 2025)	60.8	46.7	50.4 ± 0.5	28.1	**2.27**	**6.5**
YOLOv12-C2F-S2Block (ours)	**61.4**	**47.7**	**51.5 ± 0.4**	**29.3**	2.87	9.1

Table 3
CLDA-Head and C2F-S2Block are incorporated into YOLOv12 as the baseline model, upon which Spatial folding, Shuffle, and Residual structures are progressively integrated for an ablation study.

Base	Conv			Precision (%)	Recall (%)	mAP_{50}± (%)	Param. (M)	FLOPs (G)
	Spatial folding	shuffle	Residual					
backbone				**62.2**	47.7	51.8 ± 0.3	3.24	10.3
	√			61.1	47.6	51.7 ± 0.4	3.09	9.6
	√	√		61.9	48.2	52.1 ± 0.3	3.09	**9.6**
	√	√	√	61.5	**49.1**	**52.6** ± 0.3	**3.09**	9.7

Table 5
First, the C2F-StarBlock module is integrated into YOLOv12. Then, a comparative analysis of various attention mechanisms is conducted within this module.

Framework	C2F-StarBlock	Backbone(C2F/C3K2) Attention				Precision(%)	mAP_{50}±(%)
		CGA	MLCA	SE-Net	SCSA		
YOLOv12						60.8	50.4 ± 0.4
	√					60.9	50.6 ± 0.5
	√	√				60.8	51.1 ± 0.4
	√		√			61.0	51.1 ± 0.4
	√			√		61.2	51.3 ± 0.3
	√				√	**61.5**	**51.5 ± 0.4**

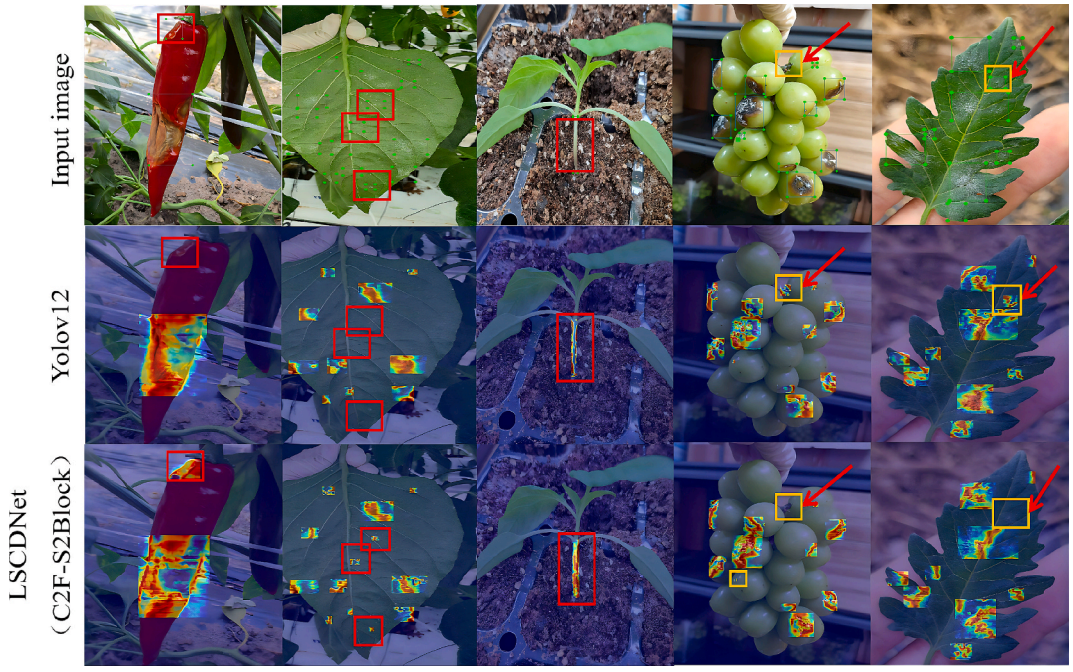

Fig. 6. Comparison of feature maps between YOLOv12 and the proposed LSCDNet(C2F-S2Block). From left to right are detection results for disease categories a7, a9, a10, a11, and a5. Red solid boxes highlight the detection performance advantages of C2F-S2Block, while yellow boxes with arrows indicate the limitations of its detection performance.

efficiently, ultimately achieving a gain where the accuracy improvement far outweighs the incremental computational cost, fully reflecting a reasonable trade-off between performance and computational expense. These results indicate that SSRConv not only enhances detection performance but also reduces computational complexity and parameter overhead, thereby facilitating a more efficient inference process.

4.5. Experiments on C2F-S2Block

To further explore the influence of C2F block designs on balancing detection accuracy and computational efficiency, this study conducts a comprehensive comparison of several YOLOv12n-C2F variants. The evaluation covers multiple metrics, including Precision, Recall, mAP_{50}, $mAP_{50\text{-}95}$, parameter count, and FLOPs, as summarized in Table 4. YOLOv12n-C2F-S2Block attains the highest values in Precision, Recall, mAP_{50}, and $mAP_{50\text{-}95}$. Compared with the YOLOv12n-C2F + C3K2 model, it improves mAP_{50} and $mAP_{50\text{-}95}$ by 1.1 % and 1.2 %, respectively, while maintaining a lightweight design with only 2.87 M parameters and 9.1 G FLOPs, thus striking a favorable balance between accuracy and efficiency. Although the YOLOv12n-C2F-DBB variant achieves comparable accuracy with an mAP_{50} of 51.1 %, it incurs a significantly higher computational cost of 11.1 G FLOPs, which may hinder practical deployment. On the other hand, the YOLOv12n-C2F-ScConv model has the smallest size, with 2.69 M parameters and 8.1 G FLOPs of computational cost, but it suffers from a notable decline in accuracy, achieving an mAP_{50} of 49.4 % and an $mAP_{50\text{-}95}$ of 27.4 %. In summary, the integration of the proposed S2Block into the C2F structure effectively enhances detection performance while keeping the model compact and computationally efficient, making it a promising solution for lightweight object detection applications.

This study conducts a comprehensive evaluation of the influence exerted by the SCSA attention mechanism on the detection performance of the C2F-S2Block module. Table 5 provides a comparative analysis of four commonly used attention mechanisms discussed in this paper: Cascaded Group Attention (CGA), Squeeze-and-Excitation Networks (SENet), Mixed Local Channel Attention (MLCA), and the proposed Spatial-Channel Synergistic Attention (SCSA). In contrast, the proposed SCSA introduces a novel synergistic attention strategy through the fusion of spatial and channel attention mechanisms, aiming to attain more precise and discriminative feature enhancement.

Furthermore, Fig. 6 presents a qualitative comparison between the proposed C2F-S2Block and the baseline YOLOv12 model. The visualizations display detection results for five representative crop disease categories: a7, a9, a10, a11, and a5. Red boxes underscore the detection advantages of C2F-S2Block by marking the underperforming regions of YOLOv12. Specifically, for cases a9 and a7, the proposed C2F-S2Block achieves notable improvements in mitigating missed detections compared with YOLOv12. Meanwhile, the heatmap of a10 highlights its

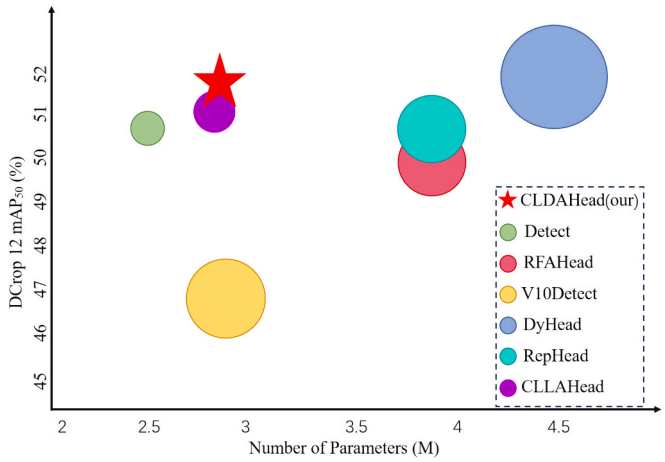

Fig. 7. The bubble chart depicts the performance of seven detection-head architectures on the DCrop12 crop-disease test set. Bubble areas are proportional to each model's FLOPs, with larger bubbles denoting greater computational overhead.

enhanced focus on disease-affected regions. Furthermore, YOLOv12's false detection in a11 is effectively alleviated. In contrast, yellow boxes indicate the limitations of C2F-S2Block in detection performance: when disease spots are extremely densely distributed or infected regions overlap, the model's localization accuracy may slightly degrade. Nevertheless, these results confirm that C2F-S2Block achieves more comprehensive and accurate detection in complex crop disease scenarios, thereby demonstrating its effectiveness in real-world agricultural applications.

4.6. Experiments on CLDA-Head

In this experiment, we selected seven representative detection heads for performance comparisons, including Detect (Owor, 2025); RFAHead (Zhang, et al., 2304), V10Detect (Wang, 2024); DyHead (Dai, 2021); RepHead (Ding, 2021); CLLAHead, and the proposed CLDA-Head. The comparison dimensions cover three aspects: model size, computational cost, and detection accuracy. As shown in Fig. 7, Although it has the smallest model size, with only 2.5 M parameters and a computational cost of 6.5 G FLOPs, the baseline Detect head achieves an mAP_{50} of just 50.4 %, which falls short of the requirements for high-precision applications. V10Detect, despite maintaining a relatively small model size of 2.9 M parameters and 8.2 G FLOPs, exhibits degraded regression performance, resulting in a reduced mAP_{50} of 46.99 %. DyHead raises mAP_{50} to 51.13 % but at the expense of a substantial 14.0 G FLOPs, compromising efficiency. The mid-complexity detection heads RepHead and CLLAHead have 2.9 M parameters, with computational costs of 7.6 and 7.2 G FLOPs, respectively. They achieve mAP_{50} scores of approximately 50.0 % and 50.67 %, demonstrating a well-balanced trade-off between accuracy and computational efficiency. In contrast, the proposed CLDA-Head achieves the highest mAP_{50} of 51.44 % with only a slight increase in computational cost, representing an improvement of approximately 1 percentage point over the baseline model. Table 6 details the mAP_{50} performance of various detection heads across small object categories. It is apparent that the CLDA-Head achieves outstanding mAP_{50} results, registering 41.4 % on a3, 47.6 % on a4, 55.5 % on a5, and 81.1 % on a10—corresponding to relative improvements of 1.3 %, 1.5 %, 1 %, and 8.1 % respectively over the Detect head. Furthermore, on all small object categories evaluated in the table, the detection accuracy of CLDA-Head consistently outperforms that of the baseline detection head Detect. The conducted results confirm that the cross-layer distribution-aware mechanism markedly enhances bounding-box regression precision and improves detection of small targets in complex backgrounds, without materially increasing model complexity.

4.7. Comparison with alternative approaches

To further validate the effectiveness of the introduced LSCDNet algorithm in crop disease detection accuracy, comparative experiments were conducted against eight mainstream disease detection methods: SSD, DETR, Mask R-CNN, DAB-DETR, DN-DETR, YOLOv8, YOLOv12, and YOLOv13. As shown in Table 7, on the DCrop12 dataset, LSCDNet achieves the highest mAP_{50} of 52.62 % with only a slight degradation in FPS, significantly outperforming YOLOv12 and other cutting-edge models. The *t*-test value between YOLOv12 and LSCDNet is 3e-6, far below the conventional 0.01 significance level. This result demonstrates that the performance discrepancy in metrics like mAP_{50} between the two models is not due to random variation but reflects a genuine and consistent difference. LSCDNet also exhibited superior performance across all categories. As shown in Fig. 8, the mAP_{50} for every class surpasses that of the baseline model. Notably, for challenging categories such as a10 and a11, LSCDNet achieved improvements of 3 % and 2 % in mAP_{50}, respectively, highlighting its robustness and effectiveness in

Table 7
Comparison of the LSCDNet model's performance with the other eight models using the testing crop disease data.

Model	Precision (%)	Recall (%)	$mAP_{50}\pm$ (%)	Param. (M)	*t*-test	FPS
SSD (Liu, et al., 2016)	51.7	43.6	48.5 ± 0.4	34	1e-7	30
DETR (Carion et al., 2020)	53.2	49.7	50.2 ± 0.6	39	3.7e-5	25
Mask-RCNN (He et al., 2017)	54.9	**52.4**	50.9 ± 0.5	31	7.2e-5	20
DAB-DETR (Liu et al., 2201)	52.6	51.9	49.8 ± 0.3	43	1e-6	21
DN-DETR (Li et al., 2022)	51.5	48.2	49.4 ± 0.5	44	4e-6	24
YOLOv8 (Varghese and Sambath, 2024)	60.3	46.3	49.7 ± 0.4	3.01	2e-6	131
YOLOv12 (Owor, 2025)	61.1	46.7	49.9 ± 0.4	2.57	3e-6	97
YOLOv13 (Tsai et al., 2025)	60.3	46.9	50.2 ± 0.3	**2.46**	4e-6	90
LSCDNet(ours)	**61.2**	49.1	**52.6 ± 0.3**	3.04	–	77

Table 6
The mAP_{50} evaluation metric results of different detection heads in the task of detecting 12 disease categories.

Model	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12
Detect	80.7	47.5	40.1	46.1	54.5	37.4	72.1	36.7	31.3	73.0	49.3	30.9
RFAHead	80.5	46.8	39.9	45.6	54.3	37.0	72.1	36.4	31.4	73.4	49.6	30.9
V10Detect	80.3	46.1	38.1	43.9	52.1	36.2	70.6	35.4	29.8	57.7	48.0	29.1
DyHead	**81.5**	47.0	40.5	46.3	55.4	37.3	**73.4**	36.7	**32.6**	79.7	**50.7**	**31.9**
RepHead	80.8	48.0	40.7	46.7	54.9	37.9	72.1	37.4	31.4	76.0	49.6	31.3
CLLAHead	80.5	48.0	40.7	46.6	54.9	37.2	72.0	37.2	31.4	80.5	49.2	31.1
CLDA-Head (ours)	81.0	**48.4**	**41.4**	**47.6**	**55.5**	**38.9**	72.9	**38.4**	32.4	**81.1**	49.6	31.8

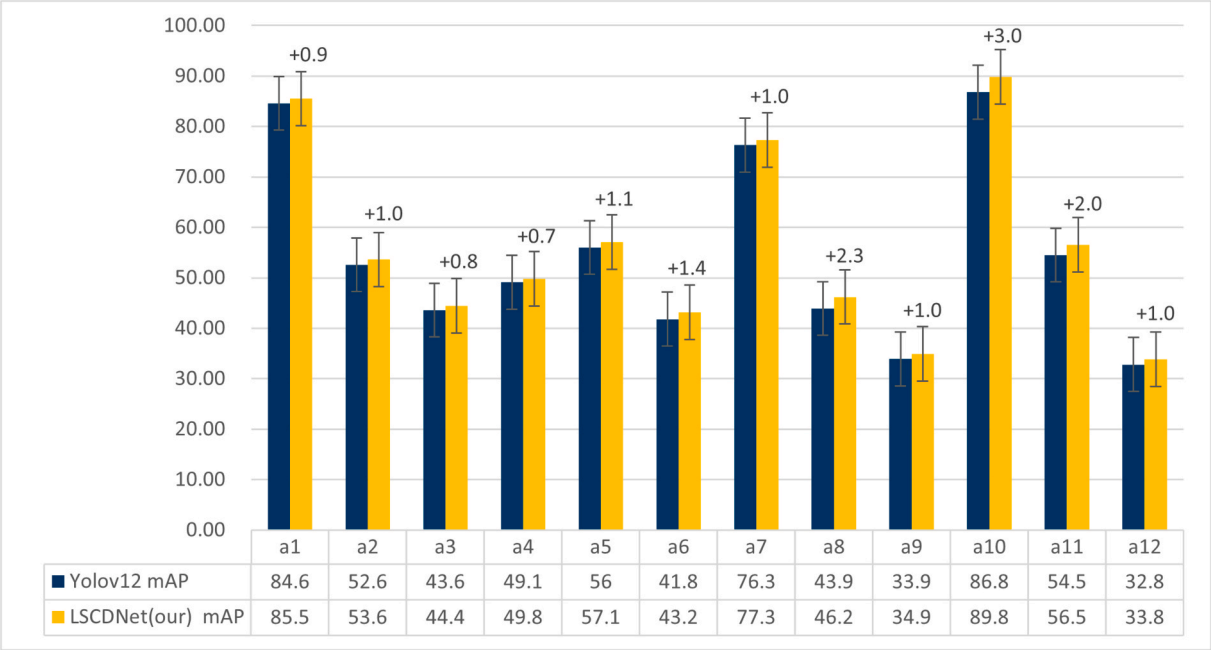

Fig. 8. Performance comparison of the baseline YOLOv12 and the proposed LSCDNet model in terms of mean Average Precision (mAP_{50}) across twelve disease categories.

multi-class detection tasks. In summary, the proposed LSCDNet model can accurately capture fine-grained features of disease-affected regions by optimizing the network architecture and enhancing feature extraction capabilities, thereby providing more reliable technical support for intelligent crop disease monitoring.

5. Conclusion

This paper proposes an automatic crop disease recognition framework based on the YOLO architecture, which is evaluated on the DCrop12 dataset. Compared with the baseline YOLOv12 model, the proposed framework introduces three core improvements: spatial information compression-channel shuffle, multi-level feature fusion-enhanced context modeling, and cross-layer attention-synergistic distribution-aware regression. Through the integration of these mechanisms, the framework achieves the synergistic enhancement of efficiency, sensitivity to fine-grained details, and multi-scale detection capability in small object detection. Compared with the YOLOv12 baseline model, the proposed model improves Precision by 0.1 percentage point, Recall by 2.4 percentage points, and the overall mAP_{50} by a notable 2.7 percentage points (a relative improvement of 5.41 %). Especially for the four typical small object disease categories a4, a9, a10, and a11, the mAP_{50} values are increased by 0.7, 1, 3, and 2 percentage points, respectively. This result effectively highlights the performance advantages of the proposed model in small object detection tasks. Although the plant disease dataset employed in the present research encompasses twelve disease categories spanning six distinct plant species, expanding the range of disease types would likely contribute to a further enhancement of the detection performance. In the future, the adaptability of the model in such complex scenarios can be further enhanced by introducing specialized datasets for extreme conditions and performing fine-tuning. Additionally, the model performance is highly dependent on large-scale and high-quality annotated data, and future work should further explore its applicability in scenarios with limited data or low-quality annotations. In terms of deployment adaptability, with only a negligible increase in the number of parameters, the proposed model achieves a substantial 2.7 percentage point improvement in detection accuracy, while its inference speed decreases only slightly to 77 frames per second (FPS). This performance highlights the model's excellent compression potential. This optimized model provides a feasible solution for low-altitude real-time crop disease detection applications. Thus, it can be inferred that implementing targeted optimizations for models of this type to strike an optimal balance between robustness and computational efficiency emerges as a key research direction for the future development of the related domain.

CRediT authorship contribution statement

Mengyao Ma: Software, Methodology, Conceptualization. **Yanfen Li:** Writing – original draft, Supervision. **Jipei Cao:** Visualization, Investigation. **Hanxiang Wang:** Writing – review & editing. **Tan N. Nguyen:** Data curation. **L. Minh Dang:** Validation, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62502271, in part by the Young Scientists Fund of the Natural Science Foundation of Shandong Province under Grant ZR2025QC630, in part by the Natural Science Foundation of Rizhao City under Grant RZ2024ZR33 and Grant RZ2024ZR34.

Data availability

Data will be made available on request.

References

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2024. Advances in medical image analysis with vision transformers: a comprehensive review.". Med. Image Anal. 91, 103000.

Carion, N., Massa, F., Synnaeve, G., et al., 2020. End-to-end object detection with transformers[C]//European conference on computer vision. Springer International Publishing, Cham, pp. 213–229.
Dai, X., et al., 2021. Dynamic head: Unifying object detection heads with attentions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, L. Zhang. Dynamic head: Unifying object detection heads with attentions. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7373-7382. 2021.
Dang, M., et al., 2024. Computer vision for plant disease recognition: a comprehensive review. Bot. Rev. 90 (3), 251–311.
Ding, X., et al., 2021. Diverse branch block: building a convolution as an inception-like unit. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Ding, X., et al., 2021. RepVgg: making VGG-style ConvNets Great again. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
K. He, G. Gkioxari, P. Dollár, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arxiv preprint arxiv:1704.04861 (2017).
Jianqiang, L., et al., 2024. Tea bud DG: a lightweight tea bud detection model based on dynamic detection head and adaptive loss function. Comput. Electron. Agric. 227, 109522.
Jogekar, R.N., Tiwari, N., 2020. A review of deep learning techniques for identification and diagnosis of plant leaf disease. In: Smart Trends in Computing and Communications: Proceedings of SmartCom 2020, pp. 435–441.
Lai, Z., Liang, G., Zhou, J., Kong, H., Yuwu, Lu., 2024. A joint learning framework for optimal feature extraction and multi-class SVM. Inf. Sci. 671, 120656.
Li, K., et al., 2020. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307.
X. Li, X. Hu, and J. Yang, Spatial group-wise enhance: Improving semantic feature learning in convolutional network, arxiv preprint arxiv:1905.09646 (2019).
F. Li, H. Zhang, S. Liu, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13619-13627.
Li, J., Wen, Y., He, L., 2023. Scconv: Spatial and channel reconstruction convolution for feature redundancy. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Li, Y., Zhou, Z., Qi, G., Gang, Hu., Zhu, Z., Huang, X., 2024. Remote sensing micro-object detection under global and local attention mechanism. Remote Sens. (Basel) 16 (4), 644.
Lin, T.-Y., et al., 2017. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
W. Liu, et al. “Ssd: Single shot multibox detector.” Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.
S. Liu, F. Li, H. Zhang, et al. Dab-detr: Dynamic anchor boxes are better queries for detr [J]. arXiv preprint arXiv:2201.12329, 2022.
Ma, Xu., et al., 2024. Rewrite the stars. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Magdy, A., et al., 2025. Lightweight faster R-CNN for object detection in optical remote sensing images. Sci. Rep. 15 (1), 16163.
M.G. Nascimento, R. Fawcett, V. A. Prisacariu, Dsconv: Efficient convolution operator. Proceedings of the IEEE/CVF international conference on computer vision. 2019.
Ouyang, D., et al., 2023. Efficient multi-scale attention module with cross-spatial learning. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Owor, N.J., et al., 2025. A unified detection pipeline for robust object detection in fisheye-based traffic surveillance. Proceedings of the IEEE/CVF International Conference on Computer Vision.
J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arxiv preprint arxiv: 1804.02767 (2018).
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.
Ross, T.-Y.-L.-P.-G., Dollár, G.K.H.P., 2017. Focal loss for dense object detection. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2980–2988.
Roy, A.M., Bhaduri, J., 2022. Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4. Comput. Electron. Agric. 193, 106694.
Roy, A.M., Bhaduri, J., 2023. DenseSPH-YOLOv5: an automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. Adv. Eng. Inf. 56, 102007.
Roy, A.M., Bose, R., Bhaduri, J., 2022. A fast accurate fine-grain object detection model based on YOLOv4 deep neural network. Neural Comput. & Applic. 34 (5), 3895–3921.
Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nat. Ecol. Evol. 3 (3), 430–439.
Shafay, M., et al., 2025. Recent advances in plant disease detection: challenges and opportunities. Plant Methods 21 (1), 140.
Y. Si, et al. “SCSA: Exploring the Synergistic Effects Between Spatial and Channel Attention. arxiv 2024.” arxiv preprint arxiv:2407.05128.
Tsai, C.-M., Li-Li, Wu., Chen, T.-Y., 2025. Enhanced fisheye object detection via yolo ensemble learning and weighted box fusion. Proceedings of the IEEE/CVF International Conference on Computer Vision.
R. Varghese, M. Sambath, Yolov8: A novel object detection algorithm with enhanced performance and robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, 2024: 1-6.
K. Vora, D. Padalia, An ensemble of convolutional neural networks to detect foliar diseases in apple plants. arXiv preprint arXiv:2210.00298 (2022).
Wang, H.e., et al., 2024. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Wang, A.o., et al., 2024. Yolov10: Real-time end-to-end object detection. Adv. Neural Inf. Proces. Syst. 37, 107984–108011.
Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534-11542. 2020.
Xie, J., Pang, Y., Nie, J., Cao, J., Han, J., 2022. Latent feature pyramid network for object detection. IEEE Trans. Multimedia 25, 2153–2163.
Xue, Z., Marculescu, R.d., 2023. Dynamic multimodal fusion. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2575–2584.
X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, X. Sun, Damo-yolo: A report on real-time object detection design. arxiv preprint arxiv:2211.15444 (2022).
Zhang, H., et al., 2021. Varifocalnet: an iou-aware dense object detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Zhang, D., et al., 2023. Detecting tomato disease types and degrees using multi-branch and destruction learning. Comput. Electron. Agric. 213, 108244.
X. Zhang, et al. “RFAConv: Innovating spatial attention and standard convolutional operation.” arXiv preprint arXiv:2304.03198 (2023).
Zhao, Y., et al., 2024. Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
H. Zhou, R. Yang, Y. Zhang, H. Duan, Y. Huang, R. Hu, X. Li, Y. Zheng, Unihead: unifying multi-perception for detection heads. IEEE Transactions on Neural Networks and Learning Systems (2024).