



AgriEyeNet: Seeing smarter in agriculture with medium-difficulty aware loss and lightweight adaptive detection

Jipei Cao¹ · Hanxiang Wang¹ · Mengyao Ma¹ · Yanfen Li¹ · Defu Qiu¹ · Yuehui Chen² · Tan N. Nguyen³ · L. Minh Dang⁴

Received: 13 July 2026 / Accepted: 8 October 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Vision-based crop disease detection is a vital field in contemporary agriculture, significantly contributing to food production and security. However, the existing object detection algorithms struggle to identify disease regions with low discrimination and irregular shapes, or the datasets with difficult and inconsistent samples, resulting in poor accuracy under complex field conditions. To overcome the above challenges, we introduce a framework incorporating a hybrid attention mechanism. The design elevates the model's feature discrimination capability for agricultural pathologies. Moreover, a slim dynamic head is designed to improve adaptability across various sizes and deformable objects. In addition, an original classification loss function is put forward to overcome common deficiency in available detection losses that leads to the under representation of both easy and moderately hard samples during the training phase. Also, this paper presents a large and challenging dataset for crop disease detection. The dataset consists of 12 categories, 136,443 images, and relevant annotation files. This research marks notable progress in intelligent agriculture and paves the way for integrating artificial intelligence technologies into farmland practices.

Keywords Geometric spatial attention · Adaptive loss function · Deformation detection head · Crop disease detection

✉ Hanxiang Wang
hanxiang@qfnu.edu.cn

✉ L. Minh Dang
minh.dlien@ou.edu.vn

Jipei Cao
jipeicv@qfnu.edu.cn

Mengyao Ma
Mmy15610723764@qfnu.edu.cn

Yanfen Li
yanfen@qfnu.edu.cn

Defu Qiu
defu.qiu@qfnu.edu.cn

Yuehui Chen
yhchen@ujn.edu.cn

Tan N. Nguyen
tnnguyen@sejong.ac.kr

¹ School of Computing Science, Qufu Normal University, Rizhao 276826, China

² Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan 250022, China

1 Introduction

As the continuous growth of the global population, there is a corresponding surge in food demand, thereby elevating the expectation for the agricultural production system to meet these escalating needs. According to the Food and Agriculture Organization of the United Nations, it is necessary to produce 50% more food by 2050 to sustain the growing global population in the context of environmental pollution and climate change [1]. Crop diseases are a key factor in declining crop and food losses, leading to a reduction in farmers' incomes. Some countries impose trade restrictions on disease-affected agricultural products, resulting in the export of agricultural products being hampered. The mentioned situation can even cause social unrest in countries or regions that depend on agriculture for their livelihoods.

³ Department of Architectural Engineering, Sejong University, Seoul 05006, Republic of Korea

⁴ Ho Chi Minh City Open University, Ho Chi Minh City 700000, Vietnam

Crop infections may stem from numerous pathogens such as fungi, bacteria, and viruses. Different types of pathogens on crops may have very similar symptoms, leading to a more complicated judgement of the diseases. Secondly, the early symptoms of many crop diseases are not obvious and closely resemble those of normally growing crops. In addition, accurate determination of crop diseases requires the use of specialized detection tools and techniques such as microscopy, nucleic acid testing as well as pathogen isolation and culture [2]. The above approaches are costly, time-consuming, and challenging, especially for the farmers without any relevant equipment available to determine the specific crop diseases. As for these issues, there is a trend towards detection aided by automated algorithms.

Machine learning has been pivotal in exploring the use of computers for detecting crop diseases. In the early stages, methods such as the Random Forest classifier were employed, achieving notable success in classifying crop diseases with a certain level of accuracy. Additionally, techniques like K-means [3] clustering were applied to group the data, ensuring the reliability of the clustering results and the ability to generalize well. However, the above methods faced limitations such as the inability to precisely locate the distribution of diseases on crops, and in some cases, they struggled with noisy data or imbalanced datasets, leading to potential overfitting issues.

With the development of deep learning (DL) technology and the improvement of hardware, an increasing number of scholars attempted to use DL for the crop diseases detection. At the beginning, most researchers tried using DL models such as Convolutional Neural Networks (CNNs). The models with only CNNs achieved high accuracy in classifying the crop diseases, demonstrating the potential of DL in this field. However, these early attempts faced challenges like requiring large amounts of labeled data and high computational power [4]. Subsequently, some scholars explored models combining CNNs with structures like Convolutional Autoencoders (CAEs) [5]. These hybrid models enhanced disease detection accuracy by lowering the input dimensions and using the results as input for the CNN. Nevertheless, they still had limitations in terms of generalization ability and adaptability to different crop species and diseases types. To address this, recent advancements have explored generalized hybrid models to improve disease classification across multiple crop species [6]. Overall, while DL shows great promise in crop disease detection, issues like data reliance, computational demands, and model interpretability still need to be addressed.

It is noteworthy that crop disease diagnosis in real-world farmland environments is fundamentally distinct from conventional image-level classification tasks [7]. Although classification models can determine whether a disease is

present, they cannot provide spatial localization information. Object detection methodologies address this limitation by delivering instance-level predictions with bounding-box annotations [8]. This approach facilitates precise disease localization, quantitative assessment of disease propagation, detailed severity evaluation, and effective discrimination of co-occurring disease instances [9]. Accordingly, despite the increased cost of instance-level annotation, we adopt a detection-oriented framework as it is better aligned with field deployment requirements.

On the basis of the above investigations, this research aims to devise a compact multi-scale attention network. This network embeds a dynamic detection head and an improved Focal Loss to tackle the challenge about detecting complex samples.

In summary, the key contributions of this study include:

1. A hybrid module combining channel and spatial attention is designed to enhance the feature extraction ability of the network.
2. A slim dynamic detection head derived from the dynamic head is proposed to make the model more convenient to deploy.
3. A loss function called Mid-Range Recovery Focal Loss (MRFL) is designed to solve the medium difficulty sample problem of Focal Loss suppression.
4. A large-scale and challenging dataset for crop disease detection is constructed, consisting of 12 categories, 136,443 images, and relevant annotation files, providing a valuable benchmark for future research in this field.

The narrative that follows is systematically arranged to facilitate the reader's comprehension of the principal components of this study. Subsequent to the overview of existing methods for crop disease detection in Section 2, the proposed methodology is elaborated in Section 3. Section 4 explains the proposed DCrop12 datasets and the conducted experimental results. Finally, Section 5 provides a comprehensive summary of this study.

2 Related work

Recently, object detection algorithms are mainly divided into one-stage detection methods and two-stage detection methods. The two-stage methods rely on region proposal network, such as R-CNN series [10–12]. Although two-stage methods provided high precision, they were limited by the computational costs. On the contrary, one-stage methods are based on bounding box regression, including YOLO [13], SSD [14], RetinaNet [15], and CenterNet [16], which significantly improved inference speed and computational efficiency. Due

to the demand for real-time processing in agricultural diseases detection, the one-stage object detection algorithms are preferred, so their corresponding feature extraction ability need to be further improved in the existing research stage. Several recent lightweight YOLO variants have jointly optimized attention modules and loss functions for crop disease detection. YOLO-CTBE embeds Triplet Attention in the backbone and designs EMASlideLoss to dynamically adjust IoU thresholds with EMA, improving adaptability to sample quality variations [17]. YOLO-DP introduces Triplet Attention and replaces the loss function with EIou to reduce position offset and shape mismatch [18]. YOLOv8-RMDA adopts inner-IoU loss to improve learning ability for small lesion samples [19]. Despite their demonstrated effectiveness, these approaches share a common limitation. The loss functions remain reliant on monotonic or dynamically adjusted weighting schemes that fail to explicitly address the specific issue of moderately difficult samples.

Feature extraction plays a crucial role in object detection, directly affecting the reliability and efficiency [20]. Various efforts have been made to enhance feature extraction in YOLO-based models. For example, based on YOLOX, the MOD-YOLO introduced the Maintaining the Original information-Deeply Separable Convolution (MODSConv) module to replace the CSP layer, decreasing the number of parameters by 19.7% while maintaining feature integrity [21]. However, the MODSConv lacked multi-scale object information, which was crucial for detecting the crop diseases at different scales. In contrast to the above model, YOLO-MSA integrated Res2Net module with different dilation rates as well as attention mechanisms to improve the multi-scale feature extraction. Although the YOLO-MSA has increased by 3.09% mAP₅₀ compared to YOLOX-S, the multi-branch feature fusion structure in YOLO-MSA results in an additional 8.05 G FLOPs contrasted with YOLOX-S, making it unsuitable for real-time applications [22]. In addition to the YOLO-based architecture, alternative detection frameworks have also been proposed to optimize feature extraction while maintaining computational efficiency. For instance, federated ensemble learning frameworks have been recently introduced to effectively segment objects across diverse multi-crop fields [23]. A CenterNet model incorporating DenseNet-77 reduces the number of parameters while maintaining accuracy, demonstrating the effectiveness of dense connections in crop disease detection [24]. Similarly, the CoffeeNet replaced traditional VGG-16 with ResNet-50, utilizing channel and spatial attention mechanisms to enhance effective feature information [25]. However, despite the stronger robustness demonstrated by CoffeeNet, a single ResNet-50 backbone cannot effectively capture multi-scale diseases features, highlighting the need for more advanced multi-scale feature fusion strategies.

Apart from CNN-based improvements, the self-attention mechanism of ViTs has been explored in crop disease detection [26]. However, standard ViT was difficult to capture fine-grained local textures and cannot fully focus on small disease areas and complex leaf shapes, as well as has more parameter quantities [27]. To address this issue, an Edge Feature Guidance module was introduced to mend the ViT's ability to capture local edge details [28]. Despite some performance improvements, ViT still introduced significant computational overhead and was still not as effective as optimized CNN based models, especially in resource constrained environments. In recent years, CNN-Transformer hybrid architectures have been introduced to tackle these constraints. A lightweight CNN-Transformer model incorporating Gray-Aware Attention and Residual Convolutional Block Attention has been developed for tomato leaf disease classification, effectively preserving fine-grained details while capturing global contextual information [29]. A lightweight fusion model combining Vision Transformer and conditional convolutional networks was designed for real-time plant disease identification with only 0.95 M parameters [30]. Despite these advances, hybrid architectures still incur substantial computational costs, and their inference latency remains a barrier for deployment on resource-constrained devices. It is a research trend to integrate self-attention into the detection head of YOLO series algorithms. Self-Attention based detection heads, such as Spatial Context Aware Module (SCAM), applied multiple branches to capture spatial and channel context information [31]. Although SCAM effectively improved small object detection, it was highly sensitive to hyperparameter selection, especially in global pooling operations. In addition, the efforts had been made to integrate multiple detection heads into YOLO based models to more effectively handle objects of various sizes [32]. However, this method significantly increased the parameter count of the model, making it prone to overfitting when trained on small datasets, thus requiring advanced regularization techniques..

Since the imbalanced data problem and the multi-scale data problem have a significant impact on detection performance, loss function optimization is also regarded as key research point. The Slide Weight Function (SWF) adjusts sample weights dynamically based on the IoU threshold and prioritized hard samples to enhance detection robustness. In order to improve the accuracy of objects, Normalized Wasserstein Distance (NWD) loss was proposed to solve the IoU offset problem, but combining NWD with SWF led to conflicts due to different optimization priorities [33]. In addition, the Focal Loss had been widely used to alleviate class imbalance by diminishing the weight of easy to classify samples [34]. Dynamic sample-weighting strategies have also been explored, where weights are adaptively determined based on sample difficulty. Dynamic Weighting Loss (DWL) adjusts

sample weights based on ground-truth bounding box area to favor tiny objects [35]. Further enhancements such as Focal-Distribution Loss introduced bounding box regularization constraints, while dynamically adjusting loss weights based on category difficulty [36]. Generalized Focal Loss (GFL) reformulates classification loss with a quality estimation branch conditioned on IoU, its modulation factor still monotonically decreases with prediction probability, leaving moderately difficult samples suppressed [37]. However, these methods typically employ binary or continuous weighting schemes that fail to explicitly object moderately difficult samples, while also introducing additional computational overhead due to complex weight adjustments.

Despite significant progress in CNN, vision transformers, and hybrid methods for the crop diseases detection, there are still challenges in balancing computational efficiency, feature extraction robustness, and loss function optimization. In order to address the existing shortcomings, this research constructs an efficient crop diseases detection method. On the one hand, a feature extraction method and a lightweight detection head are proposed, both of which are based on attention mechanism. Additionally, an adaptive classification loss is designed to expand the loss weight of moderately difficult samples.

3 Methodology

This section first introduces the overall design of the proposed AgriEyeNet model. All of the presented modules and strategies are plug and play. The proposed improvement involves the C3k2-FGN, which replaces the C3k2 module in YOLOv12 [38] (see details in Section 3.1). Inspired by the Dynamic Head [39], a Slim Dynamic Head (SDHead) is designed to enhance detection robustness in multi-scale and deformation scenes, thereby advancing its practical deployment in agricultural domains (see details in Section 3.2). Finally, the MRFL is formulated to address the concern

that Focal Loss suppresses moderately difficult samples (see details in Section 3.3). Figure 1 shows the configuration of the entire model.

3.1 FGNBlock

The standard convolution lacks the ability of global semantic integration, especially when dealing with objects in multi-scale or complex background. Although the traditional channel attention can improve the characteristic response between channels, it ignores the spatial distribution and position information of the object. In order to tackle the problem, the FGNblock is designed to enhance the representation of crucial regions and features while maintaining compactness. The FGNblock consists of three main parts, namely Fast Efficient Channel Attention (FECA), Geometric Spatial Attention (GeoSA), and Non-Linear Mixer (NLM). The structure of FGNblock and the three auxiliary modules are visualized in Fig. 2.

3.1.1 FECA

In order to improve the efficiency of information interaction between channels, a lightweight channel attention mechanism is introduced, and the simplified design is based on the ECA (Efficient Channel Attention) [40]. The formula for calculating the convolution kernel size in the original ECA module is defined by (1).

$$k = \begin{cases} \left\lfloor \frac{\log_2(C_{in}) + \alpha}{\beta} \right\rfloor + 1, & \text{if } \left\lfloor \frac{\log_2(C_{in}) + \alpha}{\beta} \right\rfloor \text{ is even,} \\ \left\lfloor \frac{\log_2(C_{in}) + \alpha}{\beta} \right\rfloor, & \text{if } \left\lfloor \frac{\log_2(C_{in}) + \alpha}{\beta} \right\rfloor \text{ is odd} \end{cases} \quad (1)$$

where C_{in} is the number of input channels, α and β mean hyperparameters.

In the FGNblock, ECA's adaptive 1D convolution kernel size is replaced by a fixed value to simplify the original ECA

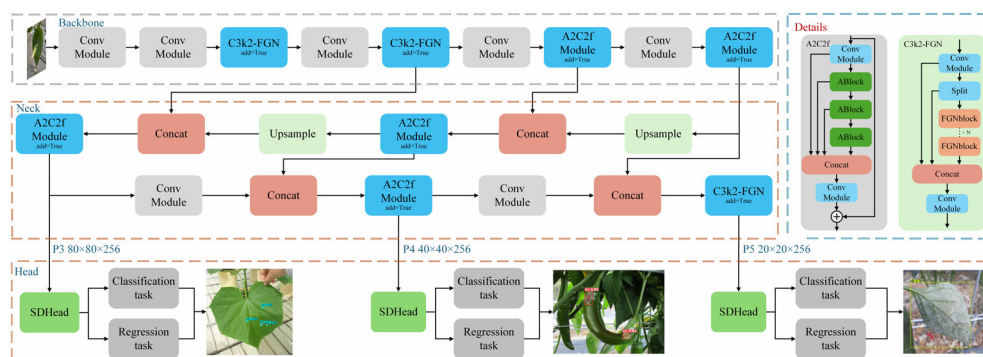


Fig. 1 The architecture design of AgriEyeNet. Note that the main improvements have been made in backbone, neck and head. Note that there are two improvements in head, one is module improvement, and the other is the classification loss in classification task

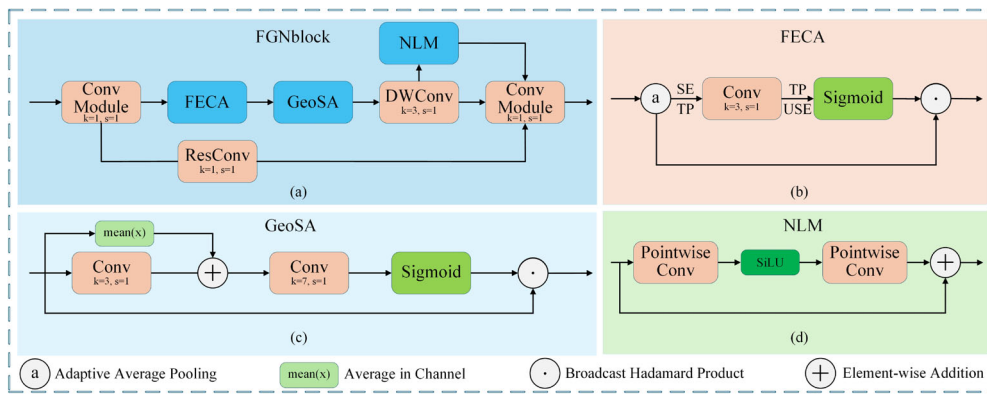


Fig. 2 The overall structure of FGNblock and its three auxiliary modules are FECA, GeoSA and NLM. SE, TP, and USE are the operations of squeeze, transpose, and contrary squeeze, respectively

module to reduce computational complexity and improve stability. This is because lightweight models have fewer convolution operation channels, and fixed convolution kernel size can reduce unnecessary operations. Firstly, the input characteristic map is adaptively averaged and pooled to obtain the importance information of each channel as shown in (2). Then the shape of the pooled result is adjusted for 1D convolution and formulated as (3).

$$Y = \text{AAP}(X) \in \mathbb{R}^{B \times C \times 1 \times 1} \quad (2)$$

$$Y = (\text{Squeeze}(Y))^T \in \mathbb{R}^{B \times 1 \times C} \quad (3)$$

where AAP is adaptive average pooling, Squeeze is to delete the specified dimension, and T represents to transpose, which is used to exchange the specified dimension. Different from the fully connected mode of the SE module [41], the 1D convolution with $k=3$ is employed to represent dependencies among neighboring channels and extract correlations across three consecutive ones (4). The FECA realizes lightweight and locally sensitive channel attention. The obtained weight information is normalized to suppress invalid channels and enhance important channels, and subsequently combined with the original feature map before being passed to the next layer (5).

$$Y'' = \text{Conv1D}_{k=3}(Y') \in \mathbb{R}^{B \times 1 \times C} \quad (4)$$

$$\text{out} = X \odot \sigma(\text{Unsqueeze}(Y''^T)) \in \mathbb{R}^{B \times C \times 1 \times 1} \quad (5)$$

where σ is the sigmoid function, Unsqueeze is used to increase dimensions, and $\odot$ is the broadcast hadamard product.

3.1.2 GeoSA

Although FECA reduces computational overhead by learning local cross-channel information through 1D convolution,

it lacks spatial awareness. The spatial dependent modeling is important for accurate object localization and edge refinement. There are some spatial attention modules (such as CBAM [42], SAM [43]) capture spatial information, but they deal with global averaging or maximum pooling separately, which will lose basic local texture details.

With the goal of strengthening spatial sensing, GeoSA utilizes a combination of global and local cues to represent features. The module captures the important location dependencies by fusing the spatial context information based on the global average and the positional response. Thus, a spatial attention map is created to modulate feature responses, striving for a balanced and precise spatial focus. The GeoSA runs on the intermediate feature graph $F \in \mathbb{R}^{B \times C \times H \times W}$. First, calculate the average of all channels to extract the rough spatial layout of global objects (Eq. (6)). At the same time, the original features are transformed nonlinearly to generate a local perception map, which is used to enhance the local change and spatial discrimination response (Eq. (7)). The above two parts are fused to form a position feature map.

$$F_{\text{mean}} = \text{Mean}(F, \text{dim} = 1) \in \mathbb{R}^{B \times 1 \times H \times W} \quad (6)$$

$$F_{\text{local}} = \text{Conv}_{3 \times 3}(F) \in \mathbb{R}^{B \times 1 \times H \times W} \quad (7)$$

In order to further refine the features and enable the model to distinguish the disease texture that is difficult to identify, the large kernel convolution is used to obtain the attention weight (Eq. (8)). The broadcast hadamard product was used between attention weight F_{spatial} and the original feature map. This process emphasizes spatial position information and suppresses irrelevant areas.

$$F_{\text{spatial}} = \sigma(\text{Conv}_{7 \times 7}(F_{\text{mean}} + F_{\text{local}})) \quad (8)$$

The GeoSA is integrated into the FGNblock, following the intermediate feature of enhancing attention. The channel attention based on FECA is supplemented by enhancing

the spatial distribution of features to ensure better alignment between the learned representation and the object position prior.

3.1.3 NLM

The Feedforward network (FFN) [44] is the core component of many deep networks, especially in the transformer-based model and the current convolution network. It helps enhance the network's nonlinear representation after attention or convolution operations. However, in convolutional backbone networks, the FFN is rarely used directly due to space constraints and large amount of computation. Inspired by the hybrid structure (such as CNN + Transformer [45], CNN + MLP [46], etc.), the NLM is introduced into the bottleneck to enrich the feature representation and interact with context information without introducing significant overhead. The function of NLM is to expand the attention obtained by FECA and GeoSA. In contrast, the NLM preserves the spatial topology and facilitates efficient pixel by pixel processing.

The NLM uses convolution instead of full connection to ensure that information is retained, which is crucial for intensive prediction tasks. The use of PWConv (Pointwise Conv) [47] makes the module lightweight and inserts it into the structure without significant additional cost.

3.2 SDHead

To elevate the detection ability in scenes with varying scales and deformations, while maintaining the highly efficient speed and computing cost, a slim detection head structure SDHead is proposed. The direct introduction of Dynamic

Head will have more parameters. Therefore, the Conv Module within the detection head is substituted by DWConv (Depthwise Conv) [47] and PWConv (Pointwise Conv), which reduces the number of inputs channels and effectively takes into account the detection precision and inference efficiency. As illustrated in Fig. 3, the SDHead uses three different attention mechanisms in sequence (9).

$$W(F) = \varphi_C(\varphi_S(\varphi_L(X) \cdot X) \cdot X) \cdot X \quad (9)$$

where φ_L is scale-sensitive attention, φ_S is spatial-sensitive attention, and φ_C is task-sensitive attention, which is applied to L , S , and C dimensions respectively. φ_L is expressed by equation (10).

$$\varphi_L(X) \cdot X = \phi(\text{Conv}_{1 \times 1}(\frac{1}{SC} \sum_{S,C} X)) \cdot X \quad (10)$$

where $\phi(\cdot)$ is the hard sigmoid function, $\phi(x) = \max(0, \min(1, (x + 1)/2))$.

Considering that S contains two dimensions, the next module needs to be divided into two stages, starting with learning deformable convolution and then aggregating the features of each level.

$$\varphi_S(X) \cdot X = \frac{1}{U} \sum_{u=1}^U \sum_{v=1}^V w_{u,v} \cdot F(u; p_v + \Delta p_v; c) \cdot \Delta m_v \quad (11)$$

where v represents the number of sparse samples, Δp_v and Δm_v correspond to the learnable system offset and modulation scalar at index v .

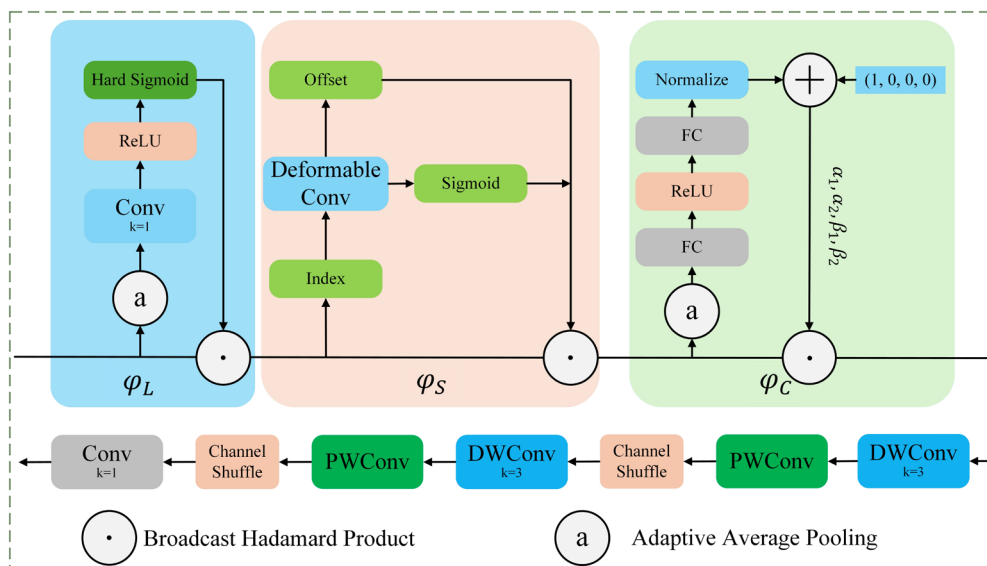


Fig. 3 The structure of SDHead module, φ_L is scale-sensitive attention, φ_S is spatial-sensitive attention, φ_C is task-sensitive attention. DWConv and PWConv are lightweight convolution methods, and Channel Shuffle is used to disrupt channels

φ_C is a task-sensitive attention function, which dynamically switches between different channels to adapt to different tasks (12).

$$\varphi_C(X) \cdot X = \max(\alpha^1(X) \cdot X_C + \beta^1(X), \alpha^2(X) \cdot X_C + \beta^2(X)) \quad (12)$$

where, $\theta(\cdot) = [\alpha^1, \alpha^2, \beta^1, \beta^2]$ is a function that learns and controls activation. It applies global average pooling on L and S dimensions to compress dimensions, after which two fully connected layers and a normalization operation are applied (see details in Fig. 3).

To minimize parameter growth from the dynamic head, the convolution module in the original detection head is replaced by DWConv and PWConv, cutting down the input channels at the same time. As shown in the lower part of Fig. 3, the standard convolution is replaced by DWConv and PWConv. After the above methods, the channel is disrupted by the Channel Shuffle module [48].

3.3 MRFL

The classification loss is fundamental to object detection accuracy in computer vision. The current object detection model used Binary Cross Entropy (BCE) loss as the classification loss function and can be expressed as equation (13).

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C [y_n \log(p_t) + (1 - y_n) \log(1 - (p_t))] \quad (13)$$

where N is the batch size, C represents the total classes, y_n corresponds to the true class of a sample and p indicates the estimated probability. The traditional BCE not only performs poorly when the categories are unbalanced, but also the penalty for misprediction increases exponentially with the probability difference. Abnormal values are easy to lead to unstable training or gradient exploding. Focal Loss is designed to solve class imbalance. As a modulation factor, $(1 - p_t)^\gamma$ can adjust the contribution of simple samples through the γ parameter, described by equations (14) and (15).

$$L_{\text{Focal}} = -\frac{1}{N} \sum_{n=1}^N [\alpha(1 - p_t)^\gamma \log(p_t)] \quad (14)$$

$$p_t = \begin{cases} p, & \text{if } y_n = 1, \\ 1 - p, & \text{others } y_n = 0 \end{cases} \quad (15)$$

The Focal Loss solved the imbalance problem between difficult and easy samples by restraining the loss from easy ones and increasing it for hard ones. However, adjusting the γ

value of the Focal Loss function will also weaken the weight of other difficulty samples, as shown in the Fig. 4. While reducing the loss of simple samples, the modulation factor also weakens the loss of other difficult samples, especially the loss of moderately difficult samples.

The total weight of BCE and Focal Loss can be expressed by equations (16) and (17).

$$I_{\text{BCE}}^{\text{all}} = \int_0^1 -\log(p_t) dp_t = [-p_t \log(p_t) + p_t]_0^1 = 1 \quad (16)$$

$$I_{\text{Focal}}^{\text{all}} = \int_0^1 (1 - p_t) \cdot (-\log(p_t)) dp_t \quad (17)$$

A comparison is to be made regarding the proportion of total weight associated with moderately difficult samples. The sample interval can be reasonably assumed to be (LP, HP). The sum of the weights of BCE and Focal Loss in the moderately difficult samples can be expressed by equations (18) and (19).

$$I_{\text{BCE}}^{\text{mid}} = (\text{LP}) \log(\text{LP}) - (\text{HP}) \log(\text{HP}) + \text{HP} - \text{LP} \quad (18)$$

$$I_{\text{Focal}}^{\text{mid}} = \int_{\text{LP}}^{\text{HP}} (1 - p_t)^\gamma \cdot (-\log(p_t)) dp_t \quad (19)$$

$I_{\text{BCE}}^{\text{all}}$ and $I_{\text{BCE}}^{\text{mid}}$ can be calculated directly through the integral formula, as well as $I_{\text{Focal}}^{\text{all}}$ and $I_{\text{Focal}}^{\text{mid}}$ can use Romberg quadrature formula to calculate the approximate value. The approximate process of Romberg quadrature formula is as follows. First, divide the integration interval (LP, HP) into several cells, and initially use the trapezoidal formula to approximate the integration value (20).

$$T(0, 0) = \frac{h}{2} (f(\text{LP}) + f(\text{HP})) \quad (20)$$

where $f(p_t) = (1 - p_t)^\gamma \cdot (-\log(p_t))$, h is the step size and $h = \text{HP} - \text{LP}$.

Then, each iteration subdivides the interval and adds new function values to obtain a new trapezoidal formula approximation (21).

$$T(k, 0) = \frac{1}{2} T(k-1, 0) + h \sum_{i=1}^{2^{k-1}} f(\text{LP} + (2i-1)h) \quad (21)$$

where the new step size is $\frac{h}{2}$. In order to accelerate convergence and improve accuracy, Richardson extrapolation method is introduced to recursively update the higher-order integral approximation (22).

$$T(k, j) = T(k, j-1) + \frac{T(k, j-1) - T(k-1, j-1)}{4^j - 1} \quad (22)$$

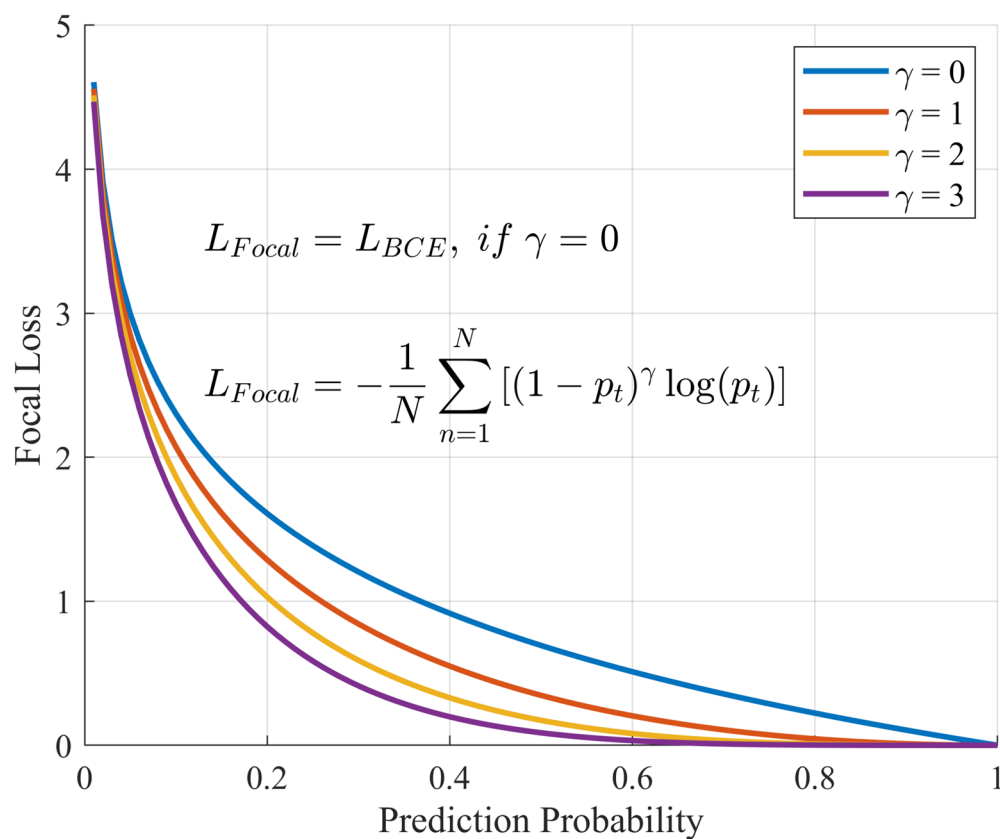


Fig. 4 The curves of different colors represent Focal Loss under different values of γ . If p_t is around 0.5, it is considered to be a moderately difficult sample

Fig. 5 The approximation of I_{Focal}^{all} with varying γ . The approximate error is less than 10^{-6}

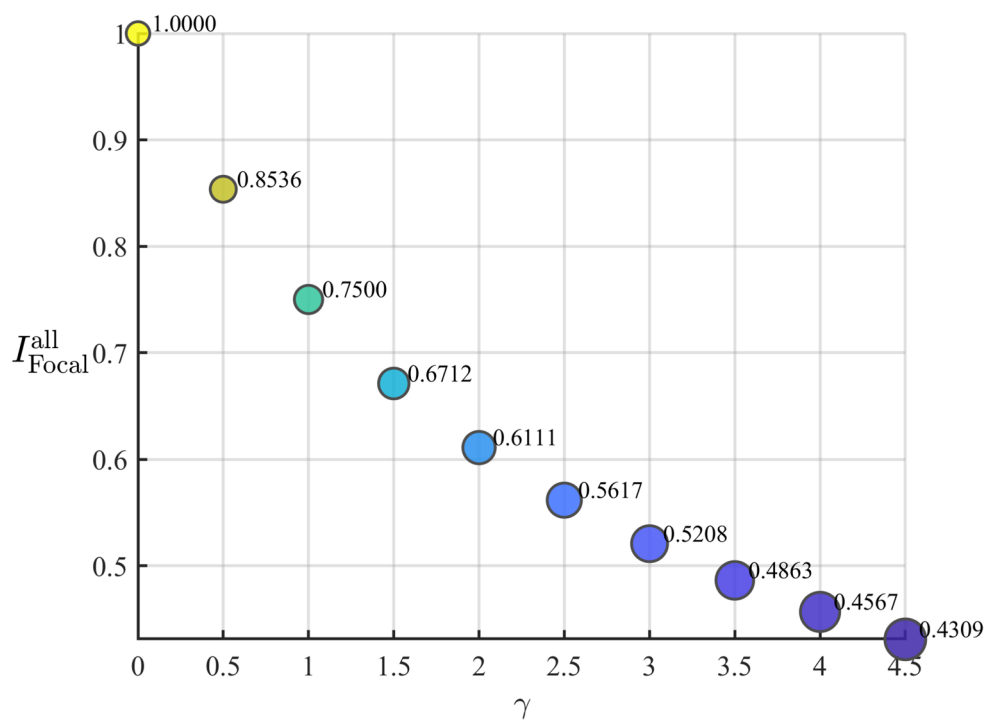


Table 1 Values of R_{BCE} and R_{Focal} in different intervals. I_{Focal}^{mid} is calculated by Romberg quadrature formula

LP	HP	I_{BCE}^{mid}	I_{Focal}^{mid}	$R_{BCE}(\%)$	$R_{Focal}(\%)$
0.1	0.9	0.6646	0.3080	66.46	50.40
0.2	0.8	0.4566	0.1680	45.66	27.49
0.3	0.7	0.2885	0.0871	28.85	14.25
0.4	0.6	0.1400	0.0368	14.00	6.02

where $j = 1, 2, \dots, k$. The iteration proceeds until the error tolerance condition $|T(k, k) - T(k - 1, k - 1)| < \epsilon$ is satisfied.

Figure 5 displays the value of I_{Focal}^{all} . In order to prove that Focal Loss weakens the proportion of moderately difficult samples, calculate $R_{BCE} = \frac{I_{BCE}^{mid}}{I_{BCE}^{all}} \times 100\%$ and $R_{Focal} = \frac{I_{Focal}^{mid}}{I_{Focal}^{all}} \times 100\%$. It can be seen from the Table 1 that Focal Loss significantly weakens the samples with a prediction probability of around 0.5 compared with BCE. The MRFL was designed to surmount the above problem. This method saves the medium difficulty samples from the suppression of Focal Loss.

MRFL controls the contribution of different difficulty samples more finely than Focal Loss. Specifically, we divided the prediction probability into three intervals and used different weight calculation methods for samples in different intervals. The expression of MRFL is described by equation (23).

$$MRFL = \begin{cases} -(\lambda - p_t)^\gamma \log(p_t), & p_t < LP, \\ -(\frac{\mu}{p_t})^\gamma \log(p_t), & LP < p_t < HP, \\ -(1 - p_t)^\gamma \log(p_t), & p_t > HP \end{cases} \quad (23)$$

where $\gamma = -\log(p_t)$, $\lambda (> 1)$ and μ are hyperparameters. The γ is determined according to the sample prediction probability of the current batch. The MRFL has a detailed description in Algorithm 1. For different datasets, these parameters need to be adjusted differently to achieve the best performance. Our original intention in designing MRFL is to reduce the inhibition of the loss of moderately difficult samples, and adjust the difficulty of simple samples and difficult samples to realize the weighting of difficult samples.

4 Experimental results

4.1 Dataset

This study utilizes data provided by the National Information Society Agency of Korea. The DCrop12 was captured by Canon EOS 60D. The dataset was collected by the Farm

Algorithm 1 Mid-Range recovery focal loss.

Require: Predictions $pred$ (logits), ground-truth labels $true \in \{0, 1\}$; hyperparameters: class-balance factor α , upper threshold $h_t = HT$, lower threshold $l_t = LT$

Ensure: Output loss $Loss$

- 1: Initialize loss function $loss_fcn \leftarrow BCEWithLogitsLoss()$
- 2: $loss \leftarrow loss_fcn(pred, true)$
- 3: $pred_prob \leftarrow \sigma(pred)$
- 4: $p_t \leftarrow true \cdot pred_prob + (1 - true) \cdot (1 - pred_prob)$
- 5: $\bar{p}_t \leftarrow \text{mean}(p_t)$
- 6: $\gamma \leftarrow -\log(\bar{p}_t + 10^{-6})$
- 7: **for** each sample j **do**
- 8: **if** $p_{t,j} > h_t$ **then**
- 9: $p_j^{high} \leftarrow (1 - p_{t,j})^\gamma$
- 10: **else if** $p_{t,j} < l_t$ **then**
- 11: $p_j^{low} \leftarrow (1.5 - p_{t,j})^\gamma$
- 12: **else** $[l_t \leq p_{t,j} \leq h_t]$
- 13: $p_j^{mid} \leftarrow \left(\frac{0.5}{p_{t,j} + 10^{-6}}\right)^\gamma$
- 14: **end if**
- 15: **end for**
- 16: $modulating_factor \leftarrow p^{high} + p^{low} + p^{mid}$
- 17: $Loss \leftarrow loss \cdot modulating_factor$
- 18: **return** $Loss$

Hannong Company Limited, Nonghyup University, Yonam College, Gyeongsangbuk-do Agricultural Research & Extension Services, and Jeonbuktechnopark. Fifteen experts from Jeonbuktech Park joined the nine-month dataset annotation task. Since the proposed dataset is large-scale, an image compression technique is used during the data collection phase, where the images are sampled with a step size of 1. The pixels of the original image are compressed from 4K to 1K. The resolution reduction from 4K to 1K inevitably discards fine-grained texture details that may be critical for identifying small diseases. This trade-off should be considered when applying the model to datasets with different resolution characteristics. In this study, the data augmentation techniques used include random perturbations of brightness, saturation, and contrast, random rotation, as well as mosaic technology. At the same time, it made feature extraction more challenging for training. The dataset called DCrop12 contained a total of 136,443 images distributed in 12 categories, as shown in Fig. 6. The DCrop12 has six crop species including chili pepper, strawberry, grape, tomato, paprika and cucumber. Also, there are four different diseases, including anthracnose, powdery mildew, botrytis cinerea and damping off. Each kind of crop is divided into two diseases. To be more specific, strawberry botrytis cinerea, strawberry powdery mildew, cucumber botrytis cinerea, cucumber powder mildew, tomato powder mildew, tomato botrytis cinerea, chili pepper anthracnose, chili pepper powder mildew, paprika powder mildew, paprika damping off, grape anthracnose and grape botrytis cinerea are sequentially labeled from a1 to a12. The specific class names and disease distribution of 12 categories are displayed in Fig. 7.

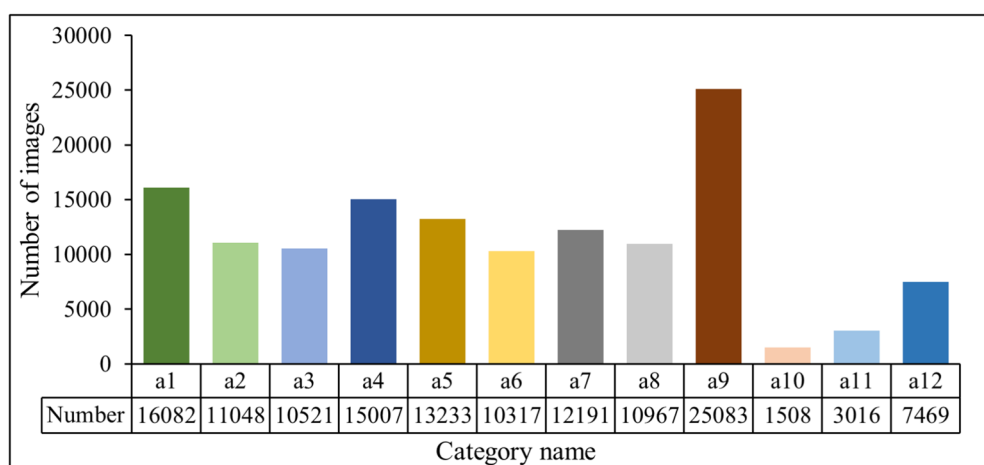


Fig. 6 Image-specific breakdown of the 12 categories in the DCrop12 dataset

4.2 Experimental environment

In terms of the experiments, we use the Ubuntu 20.04.1 operating system, which includes 256 GB of RAM, four 48 GB of NVIDIA RTX A6000 GPUs, and an Intel(R) Xeon(R) Gold 5218N processor running at 2.30 GHz. The build environment uses Python 3.11.9 and PyTorch 2.5.0. The optimizer is SGD, and the initial learning rate is set to 0.01. The training, validation, and testing of the dataset are divided in a ratio of 8:1:1. To ensure the reliability of the research findings, the

experiments were performed under the same hardware and software environments.

4.3 Analysis of C3k2-FGN

The experiment aims to validate the effectiveness of FGN-block in AgriEyeNet. The FGNblock is composed of FECA, GeoSA and NLM. For a clearer display of the network model's attention area, we visually compare our model with the YOLOv12, and embed the class activation graph



Fig. 7 The representative images selected from each category in the DCrop12 dataset

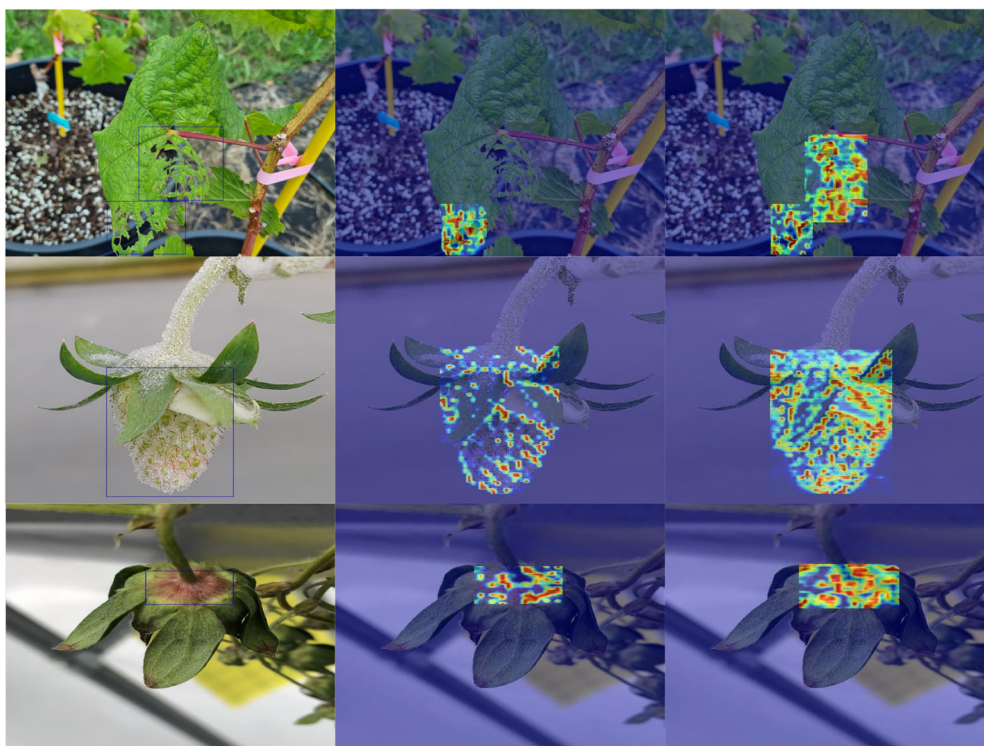


Fig. 8 The labels of original images and Grad-CAM++ results. Column one displays the original images with labels, while columns two and three show Grad-CAM++ outputs for the C3k2 and C3k2-FGN models

Grad-CAM++ behind C3k2-FGN to display the attention area of the model. As shown in the first row of Fig. 8, the C3k2-FGN pays more attention to plant disease areas, and can also identify more areas that C3k2 cannot recognize.

To determine the optimal kernel size for the fixed 1D convolution in FECA, we conducted a comparative ablation study with $k = 1, 3, 5$, and 7 . Table 2 shows that $k = 3$ strikes the optimal balance between detection accuracy and computational efficiency. The results show that $k = 3$ achieves the highest recall, mAP₅₀, mAP, and F1-score among all configurations. The performance degradation at $k = 1$ and $k = 7$ further confirms that $k = 3$ provides the most effective local cross-channel interaction for the lightweight channel attention design. Therefore, we adopt $k = 3$ as the fixed kernel size in FECA.

The FGNblock adopts serial stacking because spatial attention benefits from operating on channel-calibrated features. When channel attention is applied first, the subsequent

spatial attention can focus on the most discriminative spatial locations within a refined feature space. In contrast, parallel fusion forces spatial attention to process raw, uncalibrated features, often producing attention maps that are overly sensitive to background noise. The serial design outperforms the parallel variant across all metrics in Table 3, achieving a 1.526% improvement in mAP₅₀ and a 1.322% improvement in F1-score. This confirms that the cascaded coarse-to-fine refinement pipeline provides more effective feature enhancement than independent branch fusion, particularly for agricultural disease detection where lesion boundaries are often subtle and background interference is severe.

Table 4 provides the analysis of the experimental results of C3k2-FGN. On the basis of C3k2 module, three modules in FGNblock are added respectively to evaluate the overall performance according to precision, recall, and mAP₅₀. The highlight of C3k2-FGN is the integration of

Table 2 Ablation study of different fixed kernel sizes k in the FECA. All other settings are identical to the baseline configuration

k	Precision(%)	Recall(%)	mAP ₅₀ (%)	mAP(%)	F1-score(%)
1	60.037	46.579	50.038	26.593	52.459
3	60.178	47.619	50.423	27.992	53.167
5	61.496	46.746	50.599	27.798	53.116
7	59.342	46.872	49.580	27.913	52.375

Bold values represent the best results

Table 3 Performance comparison between serial stacking and parallel branch fusion

C3k2-FGN	Precision(%)	Recall(%)	mAP ₅₀ (%)	mAP(%)	F1-score(%)
Parallel Fusion	60.318	46.521	50.030	26.716	52.529
Serial fusion	60.924	48.250	51.556	28.849	53.851

Bold values represent the best results

multidimensional attention and nonlinear mixer enhancement while maintaining light weight, which effectively improves the ability of feature expression. According to Table 4, the baseline model achieved 61.167%, 46.348% and 49.960% precision, recall and mAP₅₀ respectively. Compared with C3k2 module, C3k2-FGN achieves 60.924% precision, 48.250% recall and 51.156% mAP₅₀. Although C3k2-FGN did not achieve the highest precision, it was 1.902% and 1.596% higher than C3k2 module in terms of recall and mAP₅₀, respectively, which was consistent with the reduction of missed detection rate of C3k2-FGN mentioned above.

The progressive improvements in Table 4 reveal a clear synergistic effect among the three components of FGNblock. When all three modules are integrated, the mAP₅₀ improvement reaches 1.60%, substantially exceeding the sum of individual gains. We attribute this super-additive effect to the complementary nature of the three modules. GeoSA provides spatial priors that guide FECA's channel selection toward disease-relevant regions, while NLM expands the representational capacity to consolidate the enhanced features from both attention modules. This cascaded design establishes a progressive refinement pipeline where each module's output serves as an enhanced input for the subsequent module.

4.4 Analysis of SDHead

To examine whether the simplification in SDHead incurs any loss in deformation modeling capability, we compare its performance with DyHead based on the results in Table 5. Although SDHead reduces parameters by 54.39% and FLOPs by 60.71%, its mAP₅₀ is only 0.061% lower than that of DyHead, while its F1-score is 0.097% higher. This marginal difference indicates that the lightweight design does not substantially degrade the model's capacity to pro-

cess deformable objects. This stems from the fact that SDHead retains the three core attention mechanisms of the original dynamic head, namely scale-sensitive attention, spatial-sensitive attention, and task-sensitive attention. The simplification only replaces standard convolutions with depthwise and pointwise convolutions and reduces input channels, which mainly affects computational cost rather than the fundamental deformable modeling logic. In particular, the spatial-sensitive attention branch still employs deformable convolution with learnable offsets and modulation scalars, preserving the essential capacity for adapting to irregular disease shapes. Therefore, SDHead strikes a desirable trade-off between efficiency and deformation modeling capability, making it more suitable for resource-constrained agricultural deployment.

4.5 Analysis of MRFL

In this part, we study the influence of different LT and HT values on MRFL. In order to find a set of parameter settings that perform well on our dataset, we designed six groups of experiments. The classification loss line chart of the training process is presented in Fig. 9, and its impact on the model is compared with the baseline model. Table 6 indicates that MRFL achieves the best performance when LT and HT are taken as 0.3 and 0.7.

The recall rate of MRFL rose to 47.028%, outperforming both the BCE baseline and standard Focal Loss. This improvement is because the baseline model tends to miss difficult samples, while MRFL focuses on moderately difficult samples and amplifies their loss. This encourages the model to reduce missed detections, thereby improving recall. In addition, some experimental groups showed a marginal drop in precision. Because over-learning of difficult samples introduces more false positives, as the boundaries of these

Table 4 The C3k2 is used as the baseline module. Comparison of C3k2-FGN with GeoSA, NLM, FECA and C3k2

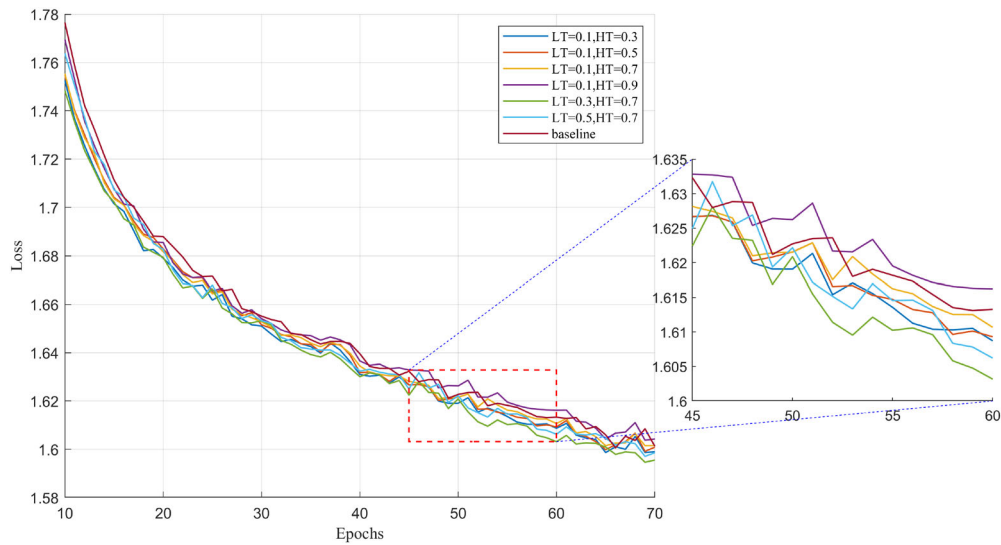
Feature Extractor	Module			Metrics (%)		
	GeoSA	NLM	FECA	Precision	Recall	mAP ₅₀
C3k2				61.167	46.348	49.960
C3k2-FGN	✓			60.178	47.619	50.423
	✓	✓		60.982	47.409	50.986
	✓	✓	✓	60.924	48.250	51.556

Bold values represent the best results

Table 5 The comparison experiment of SDHead and other detection heads, including parameters, FLOPs, mAP₅₀ and F1-score

Head	Param.(M)	FLOPs(G)	mAP ₅₀ (%)	F1-score(%)
Detect [38]	2.55	6.5	49.960	52.736
V10Detect [49]	2.98	8.2	46.999	50.338
RFAHead [50]	3.80	8.0	49.920	52.699
RepHead [51]	3.88	8.0	50.040	52.728
DyHead [39]	4.67	14.0	51.138	53.454
SDHead	2.13	5.5	51.077	53.551

Bold values represent the best results

**Fig. 9** MRFL and BCE function curves and local enlarged subgraphs of epoch from 10 to 70. With the increase of training epoch, the loss value of MRFL with LT=0.3 and HT=0.7 is the smallest**Table 6** Ablation studies of BCE, Focal Loss and MRFL with different super-parametric forms in precision, recall, mAP₅₀ and F1-score

Hyperparameter	LT	HT	Precision (%)	Recall (%)	mAP ₅₀ (%)	F1-score (%)
BCE (Baseline)	—	—	61.167	46.348	49.960	52.736
Focal Loss	—	—	60.094	46.702	50.431	52.558
MRFL Hyperparameter combinations	0.1	0.3	60.698	46.495	50.008	52.656
	0.1	0.5	60.543	46.330	49.693	52.491
	0.1	0.7	61.477	46.245	50.316	52.784
	0.1	0.9	59.700	46.782	49.992	52.457
	0.3	0.7	61.366	47.028	50.613	53.249
	0.5	0.7	61.083	46.839	50.424	53.021

Bold values represent the best results

Table 7 The comparison with other models in the DCrop12 dataset

Model	Param. (M)	FLOPs (G)	Precision (%)	Recall (%)	mAP ₅₀ (%)
SSD [53]	34	34	51.742	43.615	48.537
YOLOv5 [54]	2.51	7.2	52.013	48.372	49.360
Mask-RCNN [55]	31	178	54.914	52.468	50.926
DETR [56]	39	28	53.278	49.751	50.230
DAB-DETR [57]	43	81	52.666	51.934	49.813
Deformable-DETR [52]	41	157	52.112	53.491	50.419
DN-DETR [58]	44	94	51.598	48.264	49.423
YOLOv8 [59]	3.01	8.2	60.373	46.370	49.710
YOLOv12 [38]	2.57	6.5	61.167	46.348	49.960
YOLOv13 [60]	2.46	6.4	60.345	46.954	50.280
AgriEyeNet	2.13	5.5	61.316	49.006	52.330

Bold values represent the best results

samples are ambiguous and difficult to distinguish. Consequently, some false predictions are made with overconfident positive predictions, slightly reducing precision.

4.6 Model performance

In this section, a comprehensive comparison based on proposed DCrop12 dataset to further assess the effectiveness of AgriEyeNet. To be fair, we did not adjust the default resolution of each model to evaluate performance at the default input image size. Table 7 demonstrates that the proposed AgriEyeNet achieved the best results across multiple indicators. Collectively, the parameter quantity and calculation quantity of AgriEyeNet were only 2.13 M and 5.5 G, respectively. With fewer parameters and lower computational cost, AgriEyeNet achieved the highest precision of 61.316%. The recall rate of AgriEyeNet is 49.006%, which is 4.485% lower than the highest recall rate Deformable DETR [52]. This discrepancy can be attributed to three factors specific to agricultural disease detection. First, Deformable DETR's deformable attention enables flexible query sampling across spatial locations, which effectively covers diverse disease regions and reduces false negatives. However, this flexibility also introduces a higher number of false positives, as the attention module struggles to distinguish disease lesions from

visually similar background structures such as senescent leaf tissues or shadow patterns, resulting in lower precision. Second, crop disease lesions frequently exhibit irregular, fragmented, or faded boundaries that do not conform to the object-centric inductive bias of standard transformer attention, making precise localization challenging and reducing the IoU. Deformable DETR relies on a fixed set of object queries without explicit modeling of intra-class variability in disease morphology, which is substantial across different growth stages and pathogen strains. Our AgriEyeNet, by contrast, employs a CNN backbone with dedicated spatial-channel attention and a lightweight deformable detection head specifically tuned for agricultural pathologies, achieving better precision and a superior precision-recall balance. The mAP₅₀ of AgriEyeNet reached an impressive 52.330%, so our improvement is effective. As shown in Table 8, the incremental addition of each module yields steady gains in recall and mAP₅₀, and the complete model attains the best overall performance. This consolidated view facilitates a clear evaluation of the separate and joint contributions of C3k2-FGN, SDHead, and MRFL.

In this experiment, the images with different disease degrees were tested and evaluated against the baseline model in Fig. 10. The image describes 4 categories of crop diseases, which are annotated with bounding boxes and labels.

Table 8 Stepwise ablation of C3k2-FGN, SDHead, and MRFL on the DCrop12 dataset

Model	C3k2-FGN	SDHead	MRFL	Precision(%)	Recall(%)	mAP ₅₀ (%)
AgriEyeNet				61.167	46.348	49.960
	✓			60.924	48.250	51.556
	✓	✓		61.041	48.689	51.418
	✓	✓	✓	61.316	49.006	52.330

Bold values represent the best results

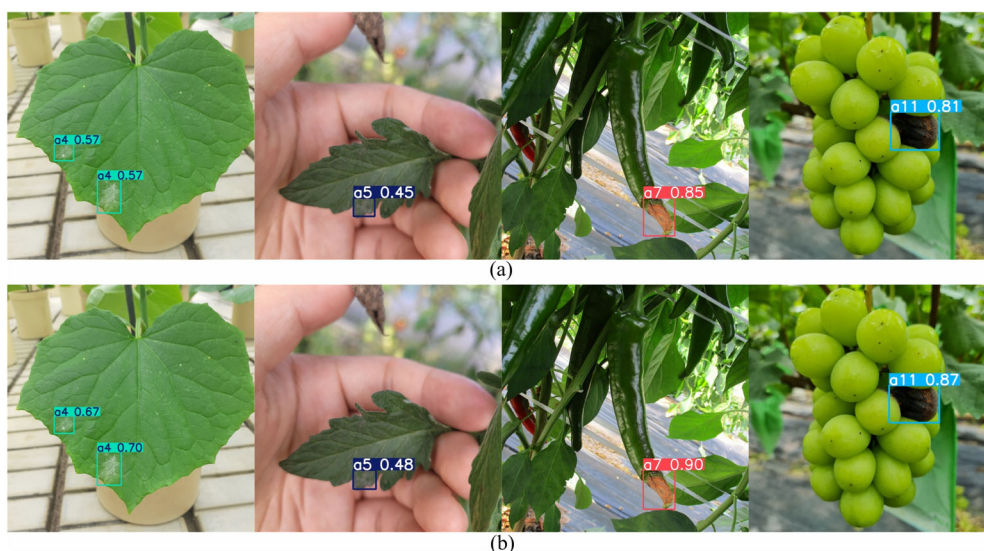


Fig. 10 The test results of YOLOv12 and AgriEyeNet are used respectively. (a) presents the results of YOLOv12. (b) presents the results of AgriEyeNet

Compared with YOLOv12, the AgriEyeNet has better performance. For the objects in the a5 class that are even difficult for human to distinguish, AgriEyeNet can gain high confidence. This shows that AgriEyeNet has good robustness and high accuracy. The robust performances are very important for deploying in the real agricultural environment.

5 Conclusion

This study presents a new detection architecture using the YOLO framework was designed to detect multiple categories of crop diseases. The proposed AgriEyeNet model adds FGNblock to address the limitation of the baseline model in accurately extracting the disease area. Since the crop disease area is irregular, AgriEyeNet introduces a slim dynamic detection head, which enhances the object detection model's effectiveness in multi-scale and deformation scenes. At the same time, an efficient classification loss function named MRFL is proposed, which alleviates the problem of traditional Focal Loss suppressing moderately difficult samples. On the whole, the AgriEyeNet obtained 52.330% mAP₅₀ on the proposed DCrop12 dataset while maintaining a less parameters and low computational overhead. In addition, our presented model is lightweight enough and easy to deploy on edge computing devices, which is meaningful for the current research on smart agriculture.

The current study has several limitations that warrant discussion. Firstly, the proposed method was evaluated on a dataset collected from South Korea, encompassing specific

crop varieties, climatic conditions, and pathogen populations that may differ from other regions. This geographical bias may limit the model's generalization to other agricultural zones, where disease symptoms can manifest differently due to variations in host cultivars, environmental stressors, and pathogen strains. To evaluate its robustness, it is necessary to conduct more extensive validation on datasets from diverse geographic origins. For future work, we plan to explore cross-regional domain adaptation through unsupervised style transfer techniques. Secondly, although improvements have been made to enhance computational efficiency, the efficient training of attention-based models remains a direction for improvement.

Acknowledgements This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62502271 and Grant 62503276, in part by the Young Scientists Fund of the Natural Science Foundation of Shandong Province under Grant ZR2025QC630 and Grant ZR2026QC0803, in part by the Natural Science Foundation of Rizhao City under Grant RZ2024ZR33 and Grant RZ2024ZR34, in part by the Open Research Fund of Shandong Key Laboratory of Ubiquitous Intelligent Computing, in part by the College Students' Innovative Entrepreneurial Training Plan Program under Grant 202510446007.

Author Contributions All authors contributed to the development and implementation of the study. Conceptualization, methodology design, validation, and visualization were performed by Jipei Cao and Hanxiang Wang. Supervision was provided by Yanfen Li, Defu Qiu and Yuehui Chen. Data collection and curation were conducted by Tan N. Nguyen and L. Minh Dang. The original draft of the manuscript was written by Jipei Cao and Mengyao Ma, all authors reviewed and edited previous versions. All authors read and approved the final manuscript.

Data Availability The datasets during the current study are not publicly available due to confidentiality constraints arising from the research

project that funded this work, but are available from the corresponding author on reasonable request.

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. IPPC Secretariat (2021) Scientific review of the impact of climate change on plant pests – a global challenge to prevent and mitigate plant pest risks in agriculture, forestry and ecosystems. Technical report, Food and Agriculture Organization of the United Nations (FAO) on behalf of the IPPC Secretariat, Rome, Italy
2. Martinelli F, Scalenghe R, Davino S, Panno S, Scuderi G, Ruisi P, Villa P, Stroppiana D, Boschetti M, Goulart LR (2015) Advanced methods of plant disease detection. a review. In: *Agronomy for sustainable development*, vol 35. pp 1–25
3. Tete TN, Kamlu S (2017) Plant disease detection using different algorithms. In: *RICE*, pp 103–106
4. Abade A, Ferreira PA, Barros Vidal F (2021) Plant diseases recognition on images using convolutional neural networks: A systematic review. *Comput Electron Agric* 185:106125
5. Bedi P, Gole P (2021) Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network. *Artif Intell Agricult* 5:90–101
6. Wadhwa D, Malik K (2026) Deep learning generalized hybrid models for multi-species crop disease classification with explainable insights. *Eng Appl Artif Intell* 164:113317
7. Anitha B, Harini S, Amirtha M, Vijaymanikandan S (2025) Smart disease severity detection system using yolo and cnn for plant health monitoring. In: *2025 Third International Conference on Emerging Applications of Material Science and Technology (ICEAMST)*. IEEE, pp 1655–1659
8. Sasmoko D, Sembiring I, Widiyari IR, Karim A (2025) Vision-based deep learning models for diagnosing tropical crop diseases and ai innovations. In: *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*. IEEE, pp 1–7
9. Ding W, Abdel-Basset M, Alrashdi I, Hawash H (2024) Next generation of computer vision for plant disease monitoring in precision agriculture: A contemporary survey, taxonomy, experiments, and future direction. *Inf Sci* 665:120338
10. Xie X, Cheng G, Wang J, Yao X, Han J (2021) Oriented r-cnn for object detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 3520–3529
11. Sasirekha R, Surya V, Nandhini P, Preethy Jemima P, Bhanushree T, Hanitha G (2025) Ensemble of fast r-cnn with bi-lstm for object detection. In: *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*. IEEE, pp 1200–1206
12. Chen Y, Wang H, Li W, Sakaridis C, Dai D, Van Gool L (2021) Scale-aware domain adaptive faster r-cnn. *Int J Comput Vision* 129(7):2223–2243
13. Sangaiah AK, Yu F-N, Lin Y-B, Shen W-C, Sharma A (2024) Uav t-yolo-rice: An enhanced tiny yolo networks for rice leaves diseases detection in paddy agronomy. *IEEE Trans Netw Sci Eng* 11(6):5201–5216
14. Sun H, Xu H, Liu B, He D, He J, Zhang H, Geng N (2021) Mean-ssd: A novel real-time detector for apple leaf diseases using improved light-weight convolutional neural networks. *Comput Electron Agric* 189:106379
15. Peng H, Li Z, Zhou Z, Shao Y (2022) Weed detection in paddy field using an improved retinanet network. *Comput Electron Agric* 199:107179
16. Dananjayan S, Tang Y, Zhuang J, Hou C, Luo S (2022) Assessment of state-of-the-art deep learning based citrus disease detection techniques using annotated optical leaf images. *Comput Electron Agricult* 193:106658
17. Fan Y, Zeng Y, Nan Y (2026) Yolo-ctbe: A multi-module enhanced network for fine-grained detection of pepper leaf diseases. *Appl Soft Comput* 116013
18. Zhou T, Wei L (2025) Yolo-dp: A detection model of fifteen common rice diseases and pests. *Sci Rep* 15(1):35968
19. Ye R, Shao G, He Y, Gao Q, Li T (2024) YOLOv8-rmda: Lightweight yolov8 network for early detection of small target diseases in tea. *Sensors* 24(9):2896
20. Zhou M, Li Y, Yang G, Wei X, Pu H, Luo J, Jia W (2025) Cofnet: Contrastive object-aware fusion using box-level masks for multi-spectral object detection. *IEEE Trans Multimedia*
21. Su P, Han H, Liu M, Yang T, Liu S (2024) Mod-yolo: Rethinking the yolo architecture at the level of feature information and applying it to crack detection. *Expert Syst Appl* 237:121346
22. Yue X, Meng L (2023) Yolo-msa: A multiscale stereoscopic attention network for empty-dish recycling robots. *IEEE Trans Instrum Meas* 72:1–14
23. Sivakumar S, Payyappilly AJ, Rajan A, Arumuga Arun R, Ram-priya RS (2025) A computer vision-based crop and weed segmentation using federated ensemble learning in diverse multi-crop fields. *Eng Appl Artif Intell* 159:111773
24. Albattah W, Nawaz M, Javed A, Masood M, Albahli S (2022) A novel deep learning method for detection and classification of plant diseases. *Comput Intell Syst* 1–18
25. Nawaz M, Nazir T, Javed A, Amin ST, Jeribi F, Tahir A (2024) Cof-feenet: A deep learning approach for coffee plant leaves diseases recognition. *Expert Syst Appl* 237:121481
26. Wang H, Nguyen T-H, Nguyen TN, Dang M (2024) Pd-tr: End-to-end plant diseases detection using a transformer. *Comput Electron Agric* 224:109123
27. Pan Y, Li Y, Yao T, Ngo C-W, Mei T (2025) Stream-vit: Learning streamlined convolutions in vision transformer. *IEEE Trans Multimedia*
28. Chang B, Wang Y, Zhao X, Li G, Yuan P (2024) A general-purpose edge-feature guidance module to enhance vision transformers for plant disease identification. *Expert Syst Appl* 237:121638
29. Zou L, Yu L, Huang X, Li Q, Hu L, Li X (2025) A hybrid cnn-transformer model for tomato leaf disease classification incorporating gray-aware attention and rs-convolutional block attention mechanisms. In: *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp 5606–5611
30. Thakur PS, Chaturvedi S, Khanna P, Sheorey T, Ojha A (2024) Real-time plant disease identification: Fusion of vision transformer and conditional convolutional network with c3gan-based data augmentation. *IEEE Trans AgriFood Electron* 2(2):576–586
31. Zhang Y, Ye M, Zhu G, Liu Y, Guo P, Yan J (2024) Ffca-yolo for small object detection in remote sensing images. *IEEE Trans Geosci Remote Sens* 62:1–15
32. Sun X, Liu K, Chen L, Cai Y, Wang H (2024) Llth-yolov5: a real-time traffic sign detection algorithm for low-light scenes. *Autom Innov* 7(1):121–137
33. Yu Z, Huang H, Chen W, Su Y, Liu Y, Wang X (2024) Yolo-facev2: A scale and occlusion aware face detector. *Pattern Recogn* 155:110714
34. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, pp 2980–2988

35. Fernando KRM, Tsokos CP (2021) Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Trans Neural Netw Learn Syst* 33(7):2940–2951
36. Chaudhry R et al (2024) Sd-yolo-awdnet: A hybrid approach for smart object detection in challenging weather for self-driving cars. *Expert Syst Appl* 256:124942
37. Li X, Lv C, Wang W, Li G, Yang L, Yang J (2022) Generalized focal loss: Towards efficient representation learning for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 45(3):3139–3153
38. Tian Y, Ye Q, Doermann D (2025) Yolov12: Attention-centric real-time object detectors. *arXiv preprint [arXiv:2502.12524](https://arxiv.org/abs/2502.12524)*
39. Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, Zhang L (2021) Dynamic head: Unifying object detection heads with attentions. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 7373–7382
40. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) 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
41. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7132–7141
42. Wang W, Tan X, Zhang P, Wang X (2022) A cbam based multi-scale transformer fusion approach for remote sensing image change detection. *IEEE J Select Topics Appl Earth Observ Remote Sens* 15:6817–6825
43. Zhang X, Shang S, Tang X, Feng J, Jiao L (2021) Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification. *IEEE Trans Geosci Remote Sens* 60:1–14
44. Rezaatofghi H, Zhu T, Kaskman R, Motlagh FT, Shi JQ, Milan A, Cremers D, Leal-Taixé L, Reid I (2021) Learn to predict sets using feed-forward neural networks. *IEEE Trans Pattern Anal Mach Intell* 44(12):9011–9025
45. Yuan F, Zhang Z, Fang Z (2023) An effective cnn and transformer complementary network for medical image segmentation. *Pattern Recogn* 136:109228
46. Desai M, Shah M (2021) An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). *Clinical eHealth* 4:1–11
47. Reis HC, Turk V (2024) Potato leaf disease detection with a novel deep learning model based on depthwise separable convolution and transformer networks. *Eng Appl Artif Intell* 133:108307
48. Fan Z, Hu W, Guo H, Liu F, Xu D (2021) Hardware and algorithm co-optimization for pointwise convolution and channel shuffle in shufflenet v2. In: *2021 IEEE International conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp 3212–3217
49. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J et al (2024) Yolov10: Real-time end-to-end object detection. *Adv Neural Inf Process Syst* 37:107984–108011
50. Zhang X, Liu C, Yang D, Song T, Ye Y, Li K, Song Y (2023) Rfaconv: Innovating spatial attention and standard convolutional operation. *arXiv preprint [arXiv:2304.03198](https://arxiv.org/abs/2304.03198)*
51. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J (2021) Repvgg: Making vgg-style convnets great again. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 13733–13742
52. Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint [arXiv:2010.04159](https://arxiv.org/abs/2010.04159)*
53. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: *Computer vision—ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14. Springer, pp 21–37
54. Wang Z, Jin L, Wang S, Xu H (2022) Apple stem/calyx real-time recognition using yolo-v5 algorithm for fruit automatic loading system. *Postharvest Biol Technol* 185:111808
55. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp 2961–2969
56. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: *European conference on computer vision*. Springer, pp 213–229
57. Liu S, Li F, Zhang H, Yang X, Qi X, Su H, Zhu J, Zhang L (2022) Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint [arXiv:2201.12329](https://arxiv.org/abs/2201.12329)*
58. Li F, Zhang H, Liu S, Guo J, Ni LM, Zhang L (2022) Dn-detr: Accelerate detr training by introducing query denoising. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 13619–13627
59. Varghese R, Sambath M (2024) Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE, pp 1–6
60. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, Ding G, Du S, Wu Z, Gao Y (2025) Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint [arXiv:2506.17733](https://arxiv.org/abs/2506.17733)*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.