

Article

Research on Maize Leaf Disease Diagnosis Based on Improved YOLO11n

Jianbin Yao *, Linyuan Li, Meijia Wang, Jingke Sun and Xinjie Xue

School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; lilinyuan0914@163.com (L.L.); mikawang1109@163.com (M.W.); sjk7318@163.com (J.S.); 18937151761@163.com (X.X.)

* Correspondence: yaojbin@163.com

Abstract

As a core global food crop, the stability of maize yield is directly related to food security. However, maize leaf diseases exhibit diverse morphologies, small initial lesions, and high similarity among different types of lesions, posing significant challenges for timely and accurate identification. Therefore, this paper proposes a maize leaf disease diagnosis model based on an improved YOLO11n. First, a C3k2_AFE module is designed, which adaptively captures rich features by operating a spatial context module and a feature refinement module in parallel. In addition, an adaptive downsampling module is adopted to enhance edge and fine-grained feature extraction. Finally, by integrating the C2PSA module with the CA-SAAttention attention mechanism, symmetric collaborative modeling between the spatial and channel domains is achieved, enhancing the model's perception of diseases. Experimental results show that the improved YOLO11n model achieves an accuracy of 89.8%, mAP@50% of 88.7%, and mAP@50-95% of 71.8%, which are 1.9%, 1.5%, and 1.4% higher than those of the baseline YOLO11n model, respectively. The number of model parameters is 2.36 MB, which is 8% lower than that of the baseline model. The corn disease diagnosis model proposed in this study effectively addresses the problem of disease detection in corn under complex environmental conditions, significantly improving the accuracy of detection and providing a reference for the application of corn leaf disease detection.

Keywords: YOLO; maize leaf diseases; object detection; lightweight

1. Introduction

As one of the most widely cultivated grains in the world, maize plays a critical role in maintaining global food security [1]. Against the backdrop of continuously expanding maize cultivation in China, the issue of diseases has emerged as a key factor constraining both yield and quality [2]. Efficient disease detection is central to ensuring maize production. Given the low efficiency and significant lag associated with traditional manual identification methods, leveraging artificial intelligence technologies for early detection of leaf diseases is of great importance for precise disease control and interrupting transmission pathways [3]. In recent years, deep learning has achieved groundbreaking progress in agricultural applications. For the task of crop disease recognition, this technology effectively overcomes the reliance on manual feature extraction inherent in traditional methods, demonstrating outstanding competitiveness in recognition accuracy and generalization capability under complex environments [4].

Received: 30 May 2026

Revised: 5 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In terms of improving recognition performance, researchers have initially focused on optimizing algorithms and developing strategies for small-sample processing, striving to accelerate model convergence under limited data conditions. To address the recognition challenges brought by small samples and complex field backgrounds, Yao et al. [5], based on VGGNet16, introduced transfer learning and attention mechanisms, and proposed the NLCBAM-VGGNet16 model. The experimental results show that the accuracy of this model in the task of identifying common diseases and pests of winter wheat reaches 97.57%, effectively improving the feature extraction and classification ability of pest and disease images, and achieving high-precision recognition. He et al. [6] proposed an improved detection algorithm called MFaster R-CNN. By incorporating batch normalization layers into the convolutional layers, the method accelerated network convergence and enhanced model generalization. In addition, they innovatively introduced a center cost function and constructed a hybrid loss function, significantly improving the model's ability to distinguish between disease spots with highly similar features. Dash et al. [7] proposed a novel classification model that integrated deep feature extraction with advanced optimization strategies. Innovatively, they introduced a support vector machine as the classification backend and employed Bayesian optimization for fine-tuning hyperparameters, thereby achieving optimal alignment between deep features and statistical learning models. Rani et al. [8] proposed a corn disease classification method that integrates attention mechanisms with few-shot learning, using a pre-trained VGG16 as the backbone network and combining it with a prototypical network for few-shot disease recognition. This approach achieved a classification accuracy of 98.25% on the corn disease dataset, effectively alleviating the problem of insufficient labeled data in agricultural scenarios.

The evolution of network architectures and the capability for multi-scale feature extraction serve as core drivers for enhancing model performance. Liao et al. [9] addressed the issue of insufficient demand for citrus fruit drop detection and quantity statistics in the assessment of variety resilience, proposing a lightweight detection model named YOLO-MECD based on YOLOv11s. This method improved upon the YOLOv11 framework by introducing the EMA attention mechanism to replace the original C2PSA module, thereby enhancing the feature expression ability and reducing the model parameters; at the same time, the CSPPC module based on partial convolution was adopted to replace the C3K2 structure to enhance the feature capture ability; in addition, the MPDIoU loss function was introduced to improve the bounding box regression accuracy and accelerate model convergence. Experimental results showed that this model achieved 84.4% Precision, 73.3% Recall, and 81.6% mAP in the citrus detection task. Nagaraju et al. [10] proposed a novel deep convolutional neural network model named NPNet-19. In comparisons with mainstream pre-trained models such as DenseNet-121 and Inception V2, NPNet-19 achieved significant improvements in classification accuracy, ranging from 1.1% to 14.52%. To address the issue of insufficient multi-scale feature recognition in corn disease detection, Yao et al. [11] improved the RT-DETR model. They first redesigned the backbone network and introduced the DAttention and SCConv modules, which enhanced the detection accuracy of the model. The mAP@0.5 reached 92.0%, while effectively reducing the model's parameter quantity. Kang et al. [12] proposed a novel detection model that combines attention mechanisms with multi-scale feature fusion. By aggregating multi-dimensional information from both local and global scales, the model achieves simultaneous coverage of small lesions and large-area diseases. Liu et al. [13] proposed a spatial multi-scale efficient real-time detector, integrating a SwConv-ResNet backbone network. By redesigning the convolution approach, they achieved lightweight architecture and introduced a novel spatial multi-scale feature fusion structure capable of capturing target features of varying sizes across scales. The model achieved 93.6% accuracy and 92% mAP@50% on the PlantVillage dataset, demonstrating significantly superior detection

performance compared to current general-purpose object detection algorithms. In addition, how to enable models to accurately identify key pathological regions under complex backgrounds has become a research hotspot, giving rise to attention mechanisms. Lv et al. [14] constructed a fully connected network and convolutional neural network model integrated with an attention mechanism. By introducing attention modules to optimize multi-view imaging spectroscopy measurement techniques, they achieved efficient and precise monitoring of southern maize rust. Qian et al. [15] proposed a novel network model combining transformers with self-attention mechanisms. By tokenizing local image region information and utilizing attention mechanisms to compute correlations between local regions, they achieved deep integration from local features to global context. This architecture adaptively weighs feature importance, enabling accurate capture of disease spot features in complex visual signals while effectively suppressing background noise. Fang et al. [16] proposed a novel deep learning architecture HCA-MFFNet, which introduces a multi-scale hard coordinate attention (HCA) mechanism and constructs a multi-feature fusion network (MFFNet) to enhance the model's ability to extract disease spot features. This architecture collaboratively extracts key weight information from two spatial dimensions, preserving subtle pathological features to the greatest extent. Masood et al. [17] proposed a deep learning framework called MaizeNet. This model is based on an improved Faster R-CNN architecture, utilizing ResNet-50 as the backbone network and incorporating a spatial-channel attention mechanism. By computing and extracting deep keypoints, it effectively focuses on lesion features while reducing interference from non-target regions, thereby achieving precise localization of diseased areas in complex visual environments. To address the issue of insufficient precision in tea leaf lesion segmentation, Hu et al. [18] proposed the MDC-U-Net3+ model based on MSFFM, multi-scale attention mechanism and CRF. The experimental results show that the mAP of this model reaches 94.92%, effectively improving the accuracy and boundary detail representation of tea leaf lesion segmentation. Regarding the problems of complex background and slow inference speed in field rice disease images, Liu et al. [19] improved the MobileViT model. By optimizing the convolutional layer and Transformer structure, the model's recognition accuracy reaches 97.25%, while also having a faster inference speed and a smaller model size, which can meet the real-time disease diagnosis requirements in the field. Sun et al. [20] proposed an improved EfficientNet model for corn leaf disease identification, enhancing the model's feature extraction capability and robustness against background interference through multi-scale feature fusion and a dual pooling strategy. The model achieved a recognition accuracy of 98.32% on the test set, significantly outperforming classical classification models such as ResNet34 and DenseNet121. Du et al. [21] addressed the issues of large-scale variations of pest and disease targets, inconsistent feature levels, and severe interference from negative samples in complex agricultural scenarios, and proposed the YOLOv7-PSAFP detection network. This method is based on YOLOv7 and introduces the progressive spatially adaptive feature pyramid (PSAFP) to enhance the multi-scale feature fusion capability, and combines the variable focal loss and loss rank mining loss function to reduce the interference of invalid negative samples on the training process. Experimental results show that this method achieves an mAP of 84.7% on the filtered PlantVillage dataset, an improvement of 2.9% compared to the YOLOv7 baseline model, verifying its effectiveness in pest and disease detection in complex agricultural scenarios. Yang et al. [22] addressed the problems of the susceptibility of corn leaf spot disease lesions to interference in complex field backgrounds and the low efficiency of traditional manual identification. They proposed an improved leaf spot disease target detection method based on real field disease leaf images. This method builds a dataset based on real field disease leaves and introduces the slender neck module and GAM attention mechanism in YOLOv8 to enhance the multi-scale feature fusion capability and focus on the lesion area.

Experimental results show that the Precision, Recall, mAP50, and mAP50-95 of the improved model are 3.79%, 4.65%, 3.56%, and 7.30% higher than those of the original YOLOv8, respectively, enabling accurate positioning and recognition of corn leaf spot disease in complex field environments. The above studies mainly focus on network structure optimization, multi-scale feature fusion, and the introduction of attention mechanisms, enhancing the model's ability to extract and express features in lesion areas, effectively improving the accuracy and robustness of disease recognition.

As model performance stabilizes, lightweight design and real-time deployment capabilities have become key to advancing technology toward field-based agricultural production. Alpsalaz et al. [23] developed a lightweight and interpretable convolutional neural network model that achieves 94.97% classification accuracy for maize leaf diseases with only 1.22 million parameters, enabling real-time inference on mobile devices. By combining transfer learning with data augmentation techniques, the model demonstrates strong generalization capability. Zeng et al. [24] proposed a Lightweight Dense Scale Network (LDSNet), which incorporates an improved dense dilated convolution block and innovatively replaces traditional feature concatenation with a "coordinate attention scale fusion" mechanism, enhancing the model's sensitivity in extracting critical disease features under complex backgrounds. While maintaining extremely low computational load, it exhibits outstanding classification performance. Li et al. [25] reshaped the backbone network of YOLOv8s by integrating the lightweight ghostNet architecture and streamlined the detection head using the C3Ghost module, significantly improving inference speed while simplifying the model architecture. Gajjar et al. [26] proposed a lightweight real-time system integrating lesion localization and classification functions. The system employs an innovative deep convolutional neural network architecture and introduces a single-stage detector to achieve rapid leaf region extraction and precise disease localization. Through extensive testing in real farmland scenarios, the system demonstrated high success rates and robustness under complex conditions such as illumination fluctuations and occlusion.

To balance lightweight design with recognition performance, Amin et al. [27] developed an end-to-end deep learning framework that fuses deep features from two pre-trained networks, EfficientNetB0 and DenseNet121, achieving efficient maize leaf disease classification while keeping model parameters controllable. Sun et al. [28] proposed a dual-channel multi-scale network model with cross-attention mechanism and spatial dimensional feature fusion. They designed a dedicated color subnetwork to extract color representations less affected by illumination interference, and constructed a texture subnetwork to capture subtle texture evolution of disease spots while enhancing the weighting of key regions. In five-fold cross-validation on a self-built dataset, the model achieved an average accuracy of 97.02%. Maginga et al. [29] developed a dual-track detection framework for heterogeneous sensor data. The first track employs a hybrid CNN-LSTM model enhanced by wavelet transform preprocessing for feature extraction and classification of volatile organic compounds, while the second track utilizes an LSTM model to capture anomalous fluctuations in ultrasonic emissions from maize plants. The model achieved an F1 score of 0.96 on the VOC classification task.

In summary, deep learning in the field of maize leaf disease recognition has achieved significant breakthroughs in both recognition accuracy and environmental adaptability through the introduction of transfer learning, multi-scale feature fusion, and various attention mechanisms. However, to achieve truly efficient disease diagnosis in complex maize field environments, the following bottlenecks remain. First, the trade-off between accuracy and model complexity continues to be a prominent issue. Although architectures such as Transformers have enhanced overall perception, their substantial computational overhead limits deployment on embedded devices. Second, challenges persist in capturing disease information and the loss incurred during feature compression. Due to factors

such as drastic illumination changes, leaf overlap and occlusion, and the diverse morphology of disease spots, existing models still have room for improvement in both the depth and breadth of feature extraction. Based on YOLO11n, a lightweight feature extraction strategy is proposed, integrating C3k2_AFE and ADown to reduce information loss during downsampling while enhancing multi-scale feature representation. Additionally, the C2PSA_CASAttention mechanism is incorporated to strengthen the interaction between local texture features and global context information, thereby improving the detection capability for subtle disease symptoms. Extensive experiments demonstrate that the proposed method achieves a superior trade-off among accuracy, detection precision, and model complexity compared to the baseline YOLO11n and several representative lightweight object detectors. Therefore, this paper presents an improved YOLO11n model for corn leaf disease detection, which enhances detection accuracy while reducing model parameters and computational complexity.

2. Materials and Methods

2.1. Baseline Model

YOLO11 is a single-stage detection algorithm that balances accuracy and speed in object detection tasks, offering distinct advantages for real-time detection applications. The network architecture of YOLO11 primarily consists of three components: the backbone, the neck, and the detection head. Building upon the YOLOv8 algorithm, it adopts a more efficient C3k2 module in place of the original C2f module, thereby improving computational efficiency. The C3k2 module is derived from the C2f module and allows the selection of either C3k or Bottleneck by setting the parameter C3k to False or True. Following the spatial pyramid pooling module, a C2PSA module incorporating a spatial attention mechanism is added to enhance the model's focusing capability. In terms of the detection head, YOLO11 incorporates depthwise separable convolutions, making the model more lightweight while improving efficiency. The architecture diagram of the YOLO11 model is shown in Figure 1.

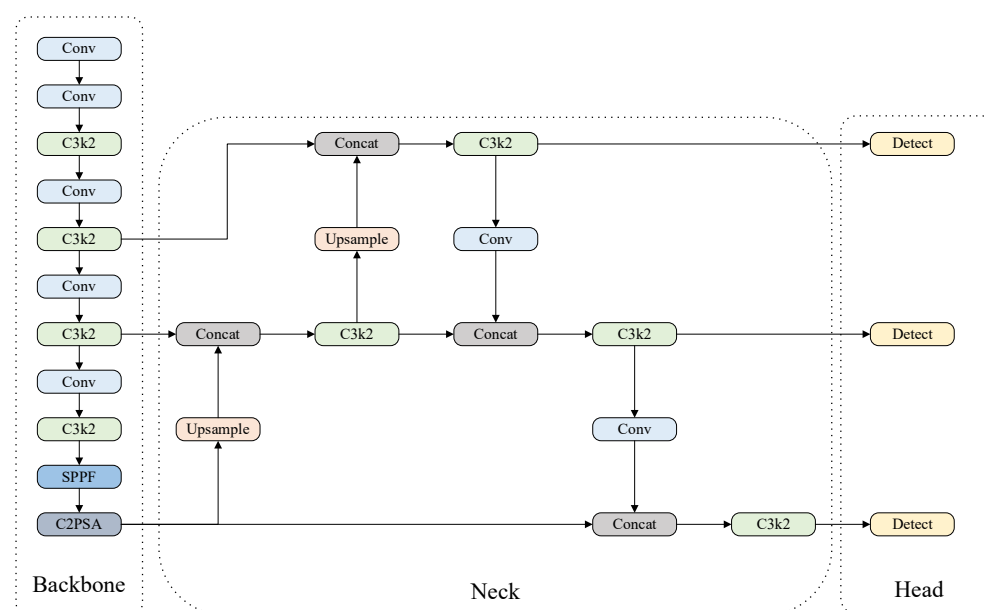


Figure 1. Module diagram of YOLO11n.

2.2. Dataset Preparation

This study utilizes the publicly available dataset Roboflow CornDiseases. The dataset comprises a total of 8310 images, covering four target categories (including ‘Maize Rust’,

‘Gray Leaf Spot’, ‘Healthy’, and ‘Leaf Blight’), as shown in Figure 2. Before training, the dataset was preprocessed using Roboflow, which included automatic orientation correction and resizing all images to 640×640 pixels. Additionally, Roboflow generated two enhanced outputs for each training image by applying random noise augmentation. The composition of the dataset is shown in Table 1.

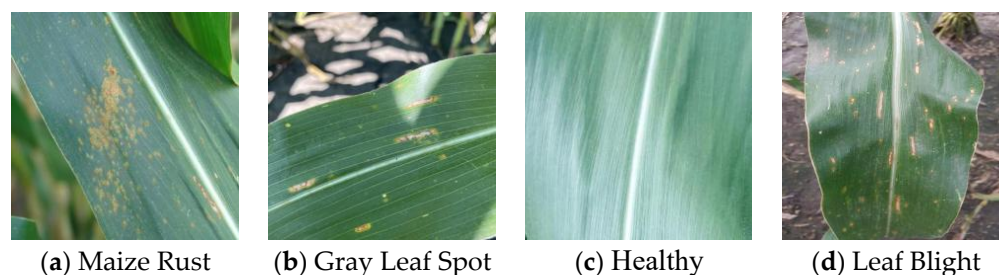


Figure 2. Maize Leaf Disease Categories.

Table 1. Dataset composition.

Category	Training Set	Validation Set
Maize Rust	2142	248
Gray Leaf Spot	2204	256
Healthy	1950	245
Leaf Blight	1092	173
Total	7388	922

2.3. Experimental Environment

Properly setting the training epochs are crucial for improving the model’s generalization ability, ensuring that the model maintains strong detection performance on data not used during training. If the training epochs are too long, the model may suffer from overfitting, while also increasing training time and computational resource consumption. Therefore, this experiment sets the number of training epochs to 200 to achieve a balance between training efficiency and model performance. Regarding the optimizer, stochastic gradient descent (SGD) demonstrates good convergence properties, so SGD is selected as the optimizer in this study. The specific training environment and parameter configurations are shown in Table 2.

Table 2. Training environment configuration.

Hyperparameter	Setting
Batch Size	64
Epochs	200
Optimizer	SGD
Image Size	640×640
Operating System	Ubuntu 22.04
Programming Language	Python 3.12
Programming Software	Pycharm 2024.3.1.1
Deep Learning Framework	PyTorch 2.5.1
GPU Model	NVIDIA GeForce RTX 4090

3. Model Design

3.1. Improved Model

For the task of maize leaf disease detection, due to the diverse morphology of disease spots and interference from complex backgrounds, the model not only requires high fea-

ture extraction precision but also must consider deployment performance on edge devices. Therefore, this paper proposes a lightweight detection scheme based on an improved YOLO11n model, aiming to balance detection accuracy, inference speed, and model parameter count.

First, to enhance the robustness of feature capture, this paper introduces an Adaptive Feature Enhancement (AFE) module to reconstruct C3k2, thereby improving the model's ability to extract multi-scale disease spot features under complex backgrounds. Second, in the downsampling stage, the ADown structure is adopted to replace traditional convolution. By employing a parallel downsampling mechanism, it effectively mitigates information loss during feature map compression, thus preserving disease spot details. Furthermore, to optimize global dependency modeling capability, this paper introduces the C2PSA_CASAttention module to replace the original C2PSA. By incorporating a convolutional additive self-attention mechanism, it strengthens the model's perception of key disease regions, effectively reducing the miss rate and improving generalization performance. The improved YOLO11n model can more accurately adapt to multi-scale lesions and complex background characteristics in maize leaves. The architecture of the improved YOLO11n model is shown in Figure 3.

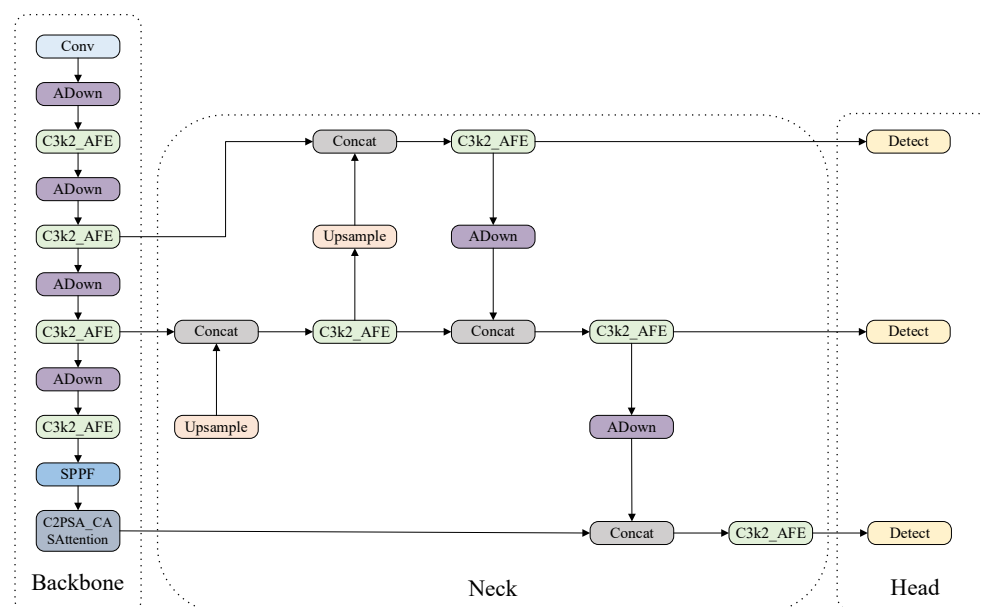


Figure 3. Improved YOLO11n module diagram.

3.2. C3k2_AFE Module

In maize leaf disease detection tasks, challenges arise from early-stage lesions that are subtle in appearance and closely resemble the color of healthy leaves. In traditional CNN architectures, feature extraction is often constrained by fixed-size convolutional kernels, making it difficult for the model to capture contextual information when dealing with cluttered backgrounds or objects with significant scale variations. Although the original C3k2 module achieved efficient feature extraction through a lightweight architecture, it mainly relied on convolution-based feature aggregation and lacked an adaptive mechanism for distinguishing informative lesion features from redundant background information. In the detection of corn leaf diseases, the diseased area typically occupies only a small portion of the image, and exhibits subtle texture changes in a complex field background. Therefore, during the feature extraction process, the discriminative features of the lesion may be weakened, thereby reducing the network's ability to accurately represent the disease pattern. To address this limitation, this paper designs the C3k2_AFE mod-

ule. The core improvement of C3k2_AFE lies in replacing the original Bottleneck with a parallel enhancement unit that integrates a Spatial Context Module (SCM) and a Feature Refinement Module (FRM).

After the input feature X enters the AFE block, it first undergoes standardized pre-processing through a Convolutional Embedding (CE) layer. Subsequently, the channel dimension is compressed to $C/2$ via a 1×1 convolution, reducing computational complexity while achieving preliminary cross-channel information integration.

$$X_{norm} = LayerNorm(X) \quad (1)$$

$$X_{ce} = Conv_{1 \times 1}(DWConv_{3 \times 3}(X_{norm}) + X_{norm}) \quad (2)$$

Among them, as the core intermediate variable X_{ce} of the module, it is fed into a dual-path enhancement branch comprising the SCM and FRM operating in parallel. This dual-path enhancement branch constitutes the core of the AFE module. On one hand, the SCM is designed to address scale variations in the scene by employing large-kernel grouped convolutions (e.g., 7×7 convolutions). Compared to stacking small-kernel convolutions, large-kernel convolutions provide a broader receptive field. The output features F_{scm} capture the spatial correlations between the target and its surrounding environment, effectively addressing the challenge of target discrimination in complex backgrounds. On the other hand, the FRM draws inspiration from sharpening filtering concepts in digital image processing. By comparing feature responses in high-frequency and low-frequency regions, FRM adaptively enhances the edge contours of targets. Particularly for semi-transparent objects or small targets with blurred boundaries, FRM extracts more discriminative semantic cues.

$$F_{fim} = Refine(X_{ce}) \otimes Attention(X_{ce}) \quad (3)$$

After obtaining the outputs of the SCM and FRM, the module does not simply perform an addition operation, but instead carries out nonlinear fusion through a convolutional multi-layer perceptron. This process achieves the “amplification” and “reconstruction” of features.

$$F_{fusion} = ConvMLP([SCM(X_{ce}), FRM(X_{ce})]) \quad (4)$$

Finally, the enhanced features are concatenated with the residual branch of C3k2, which not only preserves shallow-level detail information but also incorporates the deep semantic enhanced features processed by AFE. The structure is shown in Figure 4.

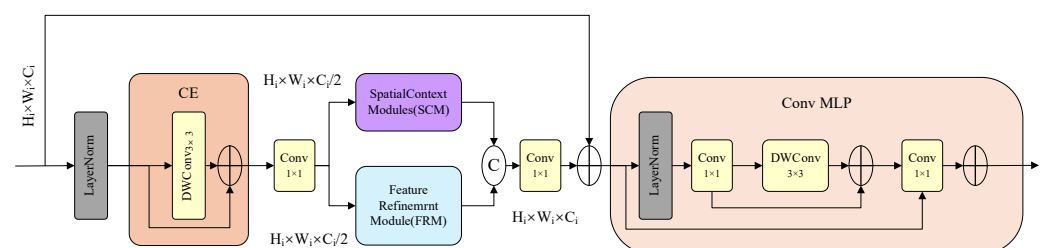


Figure 4. Adaptive Feature Enhancement (AFE) Block.

3.3. Adown Module

In object detection models, downsampling is a critical step for reducing the spatial dimensions of feature maps while capturing high-level semantic information. However, the traditional convolutional downsampling method with a stride of 2 employed by YOLO11n exhibits certain limitations when applied to the task of maize leaf disease detection. On one hand, early-stage lesions of maize diseases often present as extremely tiny, dot-like distri-

butions, and traditional convolution tends to easily lose key fine-grained details such as edges and textures during spatial dimension compression. On the other hand, the large number of parameters in traditional convolution imposes a heavy computational burden on real-time inference of the model. Therefore, in this study, an ADown downsampling module was introduced to replace the original traditional convolutional layers in the YOLO11n backbone network. During the feature compression process, more discriminative spatial details were retained. The ADown module is designed to resolve the conflict between information loss and computational redundancy during downsampling in deep convolutional neural networks. Unlike traditional strided convolution or single pooling, ADown adopts a multi-path parallel perception and feature reshaping strategy.

The ADown module achieves adaptive compression of feature space through a dynamic channel fusion mechanism. This is particularly crucial for identifying early-stage maize rust lesions, which occupy a small proportion and exhibit subtle features, as it effectively mitigates the loss of fine-grained information in deep networks and ensures the efficient transmission of lesion edge features. By incorporating depthwise separable convolution and group convolution operators, the module significantly reduces computational overhead while maintaining feature integrity, thereby striking a favorable balance between high-precision detection and real-time inference in field environments. Furthermore, ADown integrates an adaptive channel attention mechanism to achieve multi-scale deep fusion of shallow texture details and deep pathological semantics. This enables the model to precisely localize lesions even under complex illumination and background interference, markedly improving detection robustness in agricultural scenarios.

In terms of structural implementation, ADown splits the input feature map $X \in \mathbb{R}^{H \times W \times C}$ along the channel dimension into two independent sub-branches, as illustrated in Figure 5. The first is the global perception branch, which captures the statistical characteristics of the regional background through average pooling (AvgPool) and then performs a linear transformation using a 1×1 convolution to ensure smooth transmission of global semantics. The second is the local detail branch, which employs a 3×3 convolution with a specific stride to extract local edge and texture features, thereby maintaining spatial discriminability. Finally, the outputs of the two branches are fused through channel concatenation, with its mathematical expression presented as follows:

$$Y = \text{Concat} \left[\text{Conv}_{1 \times 1} \left(\text{AvgPool} \left(X_{\text{left}} \right) \right), \text{Conv}_{3 \times 3, s=2} \left(X_{\text{right}} \right) \right] \quad (5)$$

where X_{left} and X_{right} respectively represent slices of input features along the channel dimension.

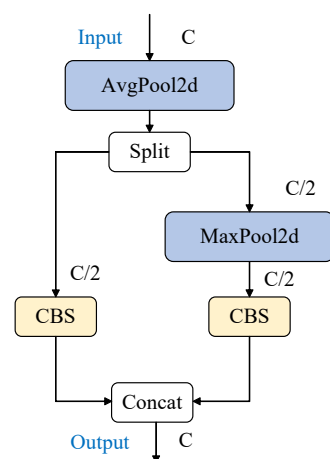


Figure 5. ADown Downsampling Module Structure Diagram.

3.4. C2PSA_CAS Attention Module

In the task of maize leaf disease detection, complex field environments often introduce severe noise interference to the model. In particular, early-stage disease lesions typically appear as dot-like distributions, and their textural features are highly similar to the physiological textures of healthy leaves. Although the original C2PSA module enhances the feature representation through PSA, its attention mechanism mainly focuses on modeling long-range dependencies and lacks sufficient convolutional inductive biases to achieve local feature extraction. In the detection of corn leaf diseases, the lesion areas usually have subtle texture changes, irregular boundaries, and small-scale spatial distribution characteristics. Therefore, the original PSA may not be able to fully capture these fine-grained local features, resulting in insufficient ability to distinguish the disease areas from the complex background. To overcome this limitation, a convolutional addition self-attention mechanism is embedded into the C2PSA module. CASAttention combines the local receptive field of convolution with the global perspective of self-attention. Through an “additive” strategy, this module effectively fuses local lesion details captured by depthwise convolution with global leaf context information obtained by self-attention. This enables the model to maintain high spatial sensitivity when confronted with clustered tiny lesions on maize leaves.

CASAttention separately extracts the spatial and channel features of $Query(Q)$ and $Key(K)$ through specific mapping functions, thereby avoiding complex matrix operations. The module primarily relies on the Convolutional Additive Token Mixer (CATM) to construct its architecture, where CATM operates on the input tensor in both the spatial and channel domains.

In the spatial domain, the module first integrates local neighborhood information using 3×3 depthwise separable convolution to establish the inductive bias of local features. Then, it compresses the feature dimensions through 1×1 convolution and generates a spatial weight map M_s using the Sigmoid activation function. In the channel domain, unlike the excessive compression in traditional SENet, CASAttention directly utilizes 1×1 convolution to achieve cross-channel information interaction and alignment, generating a channel weight map M_c . Its mathematical expression is as follows:

$$M_s = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\text{Conv}_{3 \times 3}(X))) \quad (6)$$

$$M_c = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\text{Pool}(X))) \quad (7)$$

To simultaneously utilize spatial information and channel information, CASAttention integrates spatial attention and channel attention to obtain the context mapping function $\Phi(\cdot)$, which is expressed as:

$$\Phi(X) = M_s \odot X + M_c \odot X \quad (8)$$

Unlike the traditional self-attention mechanism which calculates the correlation between Query and Key using matrix multiplication, CASAttention employs additive similarity to establish the context relationship, effectively avoiding the high computational complexity brought about by matrix multiplication and Softmax. The similarity function is defined as:

$$\text{Sim}(Q, K) = \Phi(Q) + \Phi(K) \quad (9)$$

Finally, the context information is fused using a linear mapping function, and then it is multiplied element-wise with the Value feature to obtain the output of the CASAttention. The expression is as follows:

$$O = \Gamma(\Phi(Q) + \Phi(K)) \odot V \quad (10)$$

Here, $\Gamma(\cdot) \in \mathbb{R}^N \times \mathbb{C}$ represents a linear transformation for integrating contextual information.

As illustrated in Figure 6. This mechanism is embedded into the C2PSA module of YOLO11 to form C2PSA_CASAttention. This structure retains the gradient flow advantages of Cross Stage Partial (CSP) and splits the feature map into an identity branch and an enhancement branch via the Split operation. In the enhancement branch, CASAttention replaces the original Pyramid Self-Attention (PSA), enabling the module to achieve stronger local texture recognition and global context understanding while remaining lightweight.

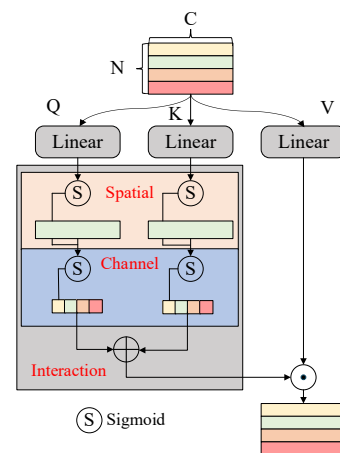


Figure 6. Conv Additive Self-Attention.

3.5. Evaluation Indicators

This paper primarily uses precision, recall, mAP, and the number of parameters as metrics to evaluate model performance.

In object detection tasks, based on the combination of ground truth and predicted labels, samples can be divided into four cases: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Precision refers to the proportion of samples predicted as positive by the model that are actually positive. It reflects the accuracy of the model's predictions. The calculation formula for precision is shown in Equation (11).

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

Recall refers to the proportion of all actual positive samples that are correctly predicted as positive by the model. It reflects the model's detection sensitivity to the target. The calculation formula for recall is shown in Equation (12).

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

mAP@50% refers to the mean average precision when the intersection over union (IoU) threshold between the predicted box and the ground truth box is 0.5. mAP@50-95% refers to the mean average precision when the IoU threshold ranges from 0.5 to 0.95 in increments of 0.05.

The calculation formulas for AP and mAP are shown in Equations (13) and (14), respectively.

$$AP = \int_0^1 PdR \quad (13)$$

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \quad (14)$$

4. Results and Discussion

4.1. Ablation Experiment

Ablation experiments are conducted to verify the effectiveness of the proposed algorithm. Modules are selectively added to the baseline model under the same dataset and experimental environment to intuitively observe the improvement effect of each module. The experimental results are shown in Table 3.

Table 3. Ablation Experiment Table.

C3k2_AFE	ADown	C2PSA_CASAttention	mAP@50 (%)	mAP@50-95(%)	Precision (%)	Parameter(M)	FLOPs (G)
			87.2	70.4	87.9	2.58	6.3
√			88.2	71.0	89.4	2.85	7.6
	√		88.4	71.1	88.1	2.10	5.1
		√	88.0	71.1	89.0	2.62	6.3
√	√		87.7	70.9	88.6	2.30	6.4
√		√	88.7	71.4	88.3	2.88	7.6
	√	√	88.6	71.4	87.8	2.14	5.2
√	√	√	88.7	71.8	89.8	2.36	6.4

Note: “√” represents that this item is adopted.

The experimental results demonstrate that, compared with the original YOLO11n model, introducing the C3k2_AFE module into the baseline model enables more sensitive capture of subtle pathological changes on maize leaves by adaptively enhancing the feature weights of key lesion regions. The ADown downsampling unit further optimizes the feature transmission process, reducing spatial information loss during downsampling by integrating pooling and convolution operations. As a result, the model achieves a 1.2% improvement in mean average precision while maintaining lightweight characteristics and effectively sustaining inference speed.

By integrating both ADown and C2PSA_CASAttention into the baseline model, the model is better able to handle complex background interference under overlapping leaves. The additive self-attention mechanism enables deep modeling of global context, allowing for more comprehensive extraction and fusion of multi-scale features. This leads to a 1.4% overall improvement in mean average precision. The synergistic effect of the three modules significantly enhances the model’s ability to disentangle maize disease features and optimizes the performance of multi-scale feature fusion.

4.2. Comparative Test

To comprehensively validate the proposed network models’ effectiveness, comparative experiments were conducted using single-stage object detection networks including YOLOv5, YOLOv8, YOLOv10, YOLO11, and RT-DETR. The experimental results are presented in Table 4.

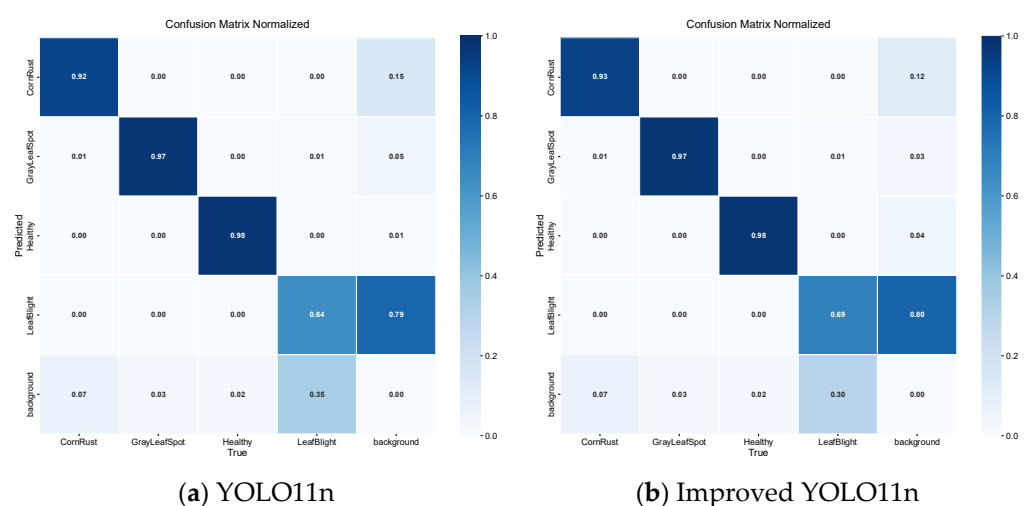
Table 4. Comparative Experiment Table.

Model	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50-95(%)	Parameter(M)
YOLOv5n	86.6	84.7	86.3	69.6	2.50
YOLOv8n	86.9	85.3	86.6	69.7	3.00
YOLOv10n	85.5	85.6	86.0	68.4	2.69
YOLO11n	87.9	86.0	87.2	70.4	2.58
YOLO12n	87.5	84.8	87.0	70.1	2.55
YOLO26n	86.9	86.1	86.5	69.7	2.38
RT-DETR	88.0	84.2	86.1	69.7	31.9
Improved YOLO11n	89.8	86.0	88.7	71.8	2.36

The experimental results show that the improved model achieves a precision of 89.8%. In comparison with other models, YOLOv5n achieves an accuracy of 86.6%, YOLOv8n achieves 86.9%, YOLOv10n achieves 85.5%, the original YOLO11n achieves 87.9%, YOLO12n achieves 87.5%, and YOLO26n achieves 86.9%. The improved model outperforms all the above comparative algorithms in terms of precision, indicating that the model has high accuracy when detecting targets. In terms of mean average precision (mAP), the improved model achieves 88.7%, which is also higher than these comparative algorithms.

4.3. Result Analysis

A confusion matrix is a structured tool for comprehensively evaluating the performance of a classification model. Its core logic lies in cross-comparing the model's predictions with the ground truth labels in the experimental environment. Each row of the matrix typically represents the true class of the samples, while each column represents the prediction results given by the model. The confusion matrix of this study is shown in Figure 7. Compared with the baseline, the proposed model shows a clearer diagonal distribution in the confusion matrix, indicating an improvement in classification consistency. Particularly, the misclassification between Leaf Blight and the background has decreased, demonstrating an enhanced ability to identify in complex scenarios.

**Figure 7.** Confusion matrix.

This describes the dynamic relationship between the model's search breadth for targets and the confidence threshold. Recall represents the proportion of all actually existing disease targets that the model successfully identifies. This curve illustrates the risk of

missed detections under different decision criteria of the model. The Recall–Confidence curve of this paper is shown in Figure 8.

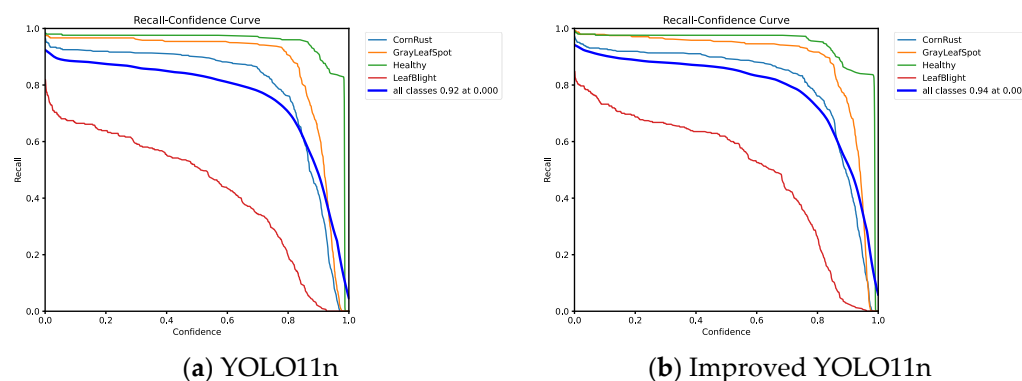


Figure 8. Recall–Confidence curve.

The Precision–Recall curve describes the relationship between precision and recall. The area under this curve directly corresponds to the most core metric in object detection—AP (Average Precision). The Precision–Recall curve of this paper is shown in Figure 9.

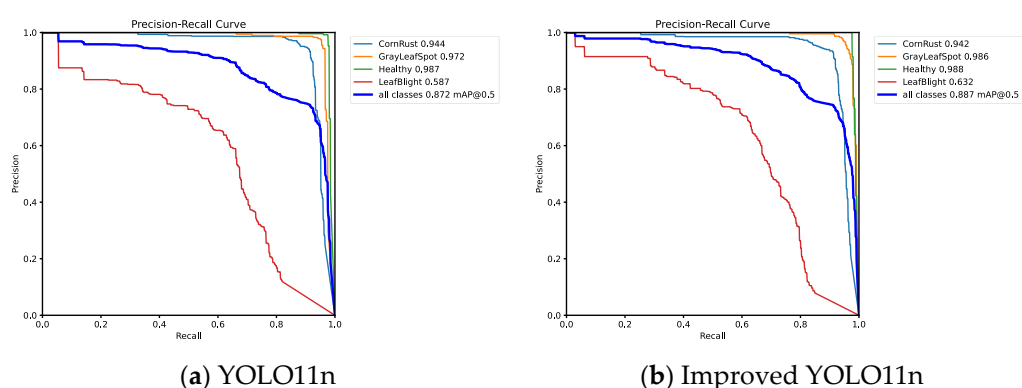


Figure 9. Precision–Recall curve.

To further clarify the performance of this study and its position in related research, it is necessary to compare it with existing studies on the identification of corn leaf diseases. Rani et al. and Sun et al. achieved high recognition accuracy in the task of corn leaf disease image classification, but their results are not suitable for direct numerical comparison with the detection performance of this study. This is because different studies have differences in data sources, category settings, image acquisition environment, sample division, model structure, and evaluation indicators; more importantly, Rani et al. and Sun et al. mainly focus on image-level classification tasks, that is, classifying disease categories for the entire leaf image, while this study uses YOLO11n for target detection, which requires simultaneously completing disease category recognition and lesion area localization, and is affected by factors such as target scale variation, background interference, leaf occlusion, and bounding box matching. Therefore, the classification Accuracy reported in existing studies and the detection indicators such as mAP used in this study reflect the model capabilities at different levels, and a lower detection indicator does not mean the model's effectiveness is insufficient in practical applications.

The position of this study in the existing literature is lightweight corn leaf disease detection research for practical agricultural monitoring needs. Different from methods that mainly aim to improve the accuracy of image-level classification, this paper aims to achieve a balance between detection accuracy, model complexity, and deployment effi-

ciency to adapt to real-time applications on resource-constrained devices. Specifically, C3k2_AFE enhances the discriminative features of lesion areas and suppresses redundant background information, improving the model's feature extraction ability in complex visual conditions; ADown retains fine-grained spatial information as much as possible during feature downsampling, thereby enhancing the perception ability for small-scale lesions; C2PSA_CASAttention combines the local inductive bias of convolution with the global modeling ability of the attention mechanism, achieving the collaborative expression of local texture details and global context information. These modules jointly enhance the model's ability to recognize and locate different-scale disease targets, enabling the proposed method to provide a solution with practical application potential for real-time detection and edge-end deployment of corn leaf diseases.

5. Conclusions

This paper proposes an improved multi-scale YOLO11n model for maize leaf disease detection to address challenges such as the diversity and similarity of disease features, large-scale variations, and field background interference.

- (1) By combining the C3k2 module with the AFE module, we propose C3k2_AFE to enhance the extraction capability of multi-scale lesion features.
- (2) The ADown downsampling module is introduced into the model's backbone network to replace ordinary convolutional blocks, significantly reducing the model's parameter count and computational load.
- (3) To reduce computational redundancy while strengthening the correlation between spatial and channel dimensions, we propose C2PSA_CASAttention. This improvement enables the model to precisely focus on lesion regions and suppress background noise, effectively reducing the missed detection rate and false detection rate in densely planted environments.

In summary, the improved YOLO11n model proposed in this paper exhibits stronger detection performance and greater advantages compared to the original YOLO11n model. The tasks that need to be carried out in the future are to conduct further research on model improvement and optimization, or to study knowledge distillation techniques, in order to transfer knowledge from large models to lightweight detectors, thereby achieving better deployment on edge devices with limited computing resources, and protecting crops from disease stress.

Author Contributions: J.Y. designed the experiments, reviewed and supervised the writing of the paper, and provided funding, L.L. performed the experiments and wrote the paper, M.W. assisted in the experiments and format revision, and J.S. and X.X. provided the datasets. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the Key scientific and technological projects of Henan Province(No.262102211046) and the North China University of Water Conservancy and Electric Power High-level Experts Scientific Research Foundation (202401014).

Data Availability Statement: All data generated or appearing in this study are available upon request by contact with the corresponding author. Furthermore, the models and code used during the study cannot be shared at this time as the data also form part of an ongoing study.

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

References

1. Yang, X.; Zhao, S.; Liu, B.; Gao, Y.; Hu, C.; Li, W.; Yang, Y.; Li, G.; Wang, L.; Yang, X.; et al. Bt maize can provide non-chemical pest control and enhance food safety in China. *Plant Biotechnol. J.* **2023**, *21*, 391–404.

2. Fan, X.; Zhou, J.; Xu, Y.; Peng, X. Corn Disease Recognition under Complicated Background Based on Improved Convolutional Neural Network. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 210–217.
3. Sujatha, R.; Chatterjee, J.M.; Jhanjhi, N.; Brohi, S.N. Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocess. Microsyst.* **2021**, *80*, 103615.
4. Rashid, R.; Aslam, W.; Aziz, R.; Aldehim, G. An early and smart detection of corn plant leaf diseases using IoT and deep learning multi-models. *IEEE Access* **2024**, *12*, 23149–23162.
5. Yao, J.; Liu, J.; Zhang, Y.; Wang, H. Identification of winter wheat pests and diseases based on improved convolutional neural network. *Open Life Sci.* **2023**, *18*, 20220632.
6. He, J.; Liu, T.; Li, L.; Hu, Y.; Zhou, G. MFaster r-CNN for maize leaf diseases detection based on machine vision. *Arab. J. Sci. Eng.* **2023**, *48*, 1437–1449.
7. Dash, A.; Sethy, P.K.; Behera, S.K. Maize disease identification based on optimized support vector machine using deep feature of DenseNet201. *J. Agric. Food Res.* **2023**, *14*, 100824.
8. Rani, R.; Sahoo, J.; Bellamkonda, S.; Kumar, S. Attention-enhanced corn disease diagnosis using few-shot learning and VGG16. *MethodsX* **2025**, *14*, 103172.
9. Liao, Y.; Li, L.; Xiao, H.; Xu, F.; Shan, B.; Yin, H. YOLO-MECD: Citrus detection algorithm based on YOLOv11. *Agronomy* **2025**, *15*, 687.
10. Nagaraju, M.; Chawla, P. Maize crop disease detection using NPNNet-19 convolutional neural network. *Neural Comput. Appl.* **2023**, *35*, 3075–3099.
11. Yao, J.; Zhu, Z.; Yuan, M.; Li, L.; Wang, M. The Detection of Maize Leaf Disease Based on an Improved Real-Time Detection Transformer Model. *Symmetry* **2025**, *17*, 808.
12. Kang, J.; Zhang, W.; Xia, Y.; Liu, W. A study on maize leaf pest and disease detection model based on attention and multi-scale features. *Appl. Sci.* **2023**, *13*, 10441.
13. Liu, Y.; Yan, F.; Li, S.; Liu, L.; Liu, Y. SMSERT-DETR: Spatial multi-scale efficient real-time detector for maize disease detection: Y. Liu et al. *J. Supercomput.* **2025**, *81*, 880.
14. Lv, Z.; Xu, B.; Zhong, L.; Chen, G.; Huang, Z.; Sun, R.; Huang, W.; Zhao, F.; Meng, R. Improved monitoring of southern corn rust using UAV-based multi-view imagery and an attention-based deep learning method. *Comput. Electron. Agric.* **2024**, *224*, 109232.
15. Qian, X.; Zhang, C.; Chen, L.; Li, K. Deep learning-based identification of maize leaf diseases is improved by an attention mechanism: Self-attention. *Front. Plant Sci.* **2022**, *13*, 864486.
16. Fang, S.; Wang, Y.; Zhou, G.; Chen, A.; Cai, W.; Wang, Q.; Hu, Y.; Li, L. Multi-channel feature fusion networks with hard coordinate attention mechanism for maize disease identification under complex backgrounds. *Comput. Electron. Agric.* **2022**, *203*, 107486.
17. Masood, M.; Nawaz, M.; Nazir, T.; Javed, A.; Alkanhel, R.; Elmannai, H.; Dhahbi, S.; Bourouis, S. MaizeNet: A deep learning approach for effective recognition of maize plant leaf diseases. *IEEE Access* **2023**, *11*, 52862–52876.
18. Hu, Y.; Guan, F.; Xie, D.; Ma, P.; Yu, Y.; Zhou, J.; Nie, Y.; Huang, L. Tea Leaf Disease Diagnosis Based on Improved Lightweight U-Net3+. *Smart Agric.* **2026**, *8*, 15–27.
19. Liu, X.; Wu, Q.; Sun, C.; Qi, C.; Zhang, G.; Lei, T.; Liang, W. Rice Disease Identification Method Based on Improved MobileViT Model and System Development. *Smart Agric.* **2026**, *8*, 28–39.
20. Sun, X.; Huo, H. Corn leaf disease recognition based on improved EfficientNet. *IET Image Process.* **2025**, *19*, e13288.
21. Du, L.; Zhu, J.; Liu, M.; Wang, L. YOLOv7-PSAFP: Crop pest and disease detection based on improved YOLOv7. *IET Image Process.* **2025**, *19*, e13304.
22. Yang, S.; Yao, J.; Teng, G. Corn leaf spot disease recognition based on improved YOLOv8. *Agriculture* **2024**, *14*, 666.
23. Alpsalaz, F.; Özüpak, Y.; Aslan, E.; Uzel, H. Classification of maize leaf diseases with deep learning: Performance evaluation of the proposed model and use of explicable artificial intelligence. *Chemom. Intell. Lab. Syst.* **2025**, *262*, 105412.
24. Zeng, W.; Li, H.; Hu, G.; Liang, D. Lightweight dense-scale network (LDSNet) for corn leaf disease identification. *Comput. Electron. Agric.* **2022**, *197*, 106943.
25. Li, R.; Li, Y.; Qin, W.; Abbas, A.; Li, S.; Ji, R.; Wu, Y.; He, Y.; Yang, J. Lightweight network for corn leaf disease identification based on improved YOLO v8s. *Agriculture* **2024**, *14*, 220.
26. Gajjar, R.; Gajjar, N.; Thakor, V.J.; Patel, N.P.; Ruparelia, S. Real-time detection and identification of plant leaf diseases using convolutional neural networks on an embedded platform. *Vis. Comput.* **2022**, *38*, 2923–2938.

27. Amin, H.; Darwish, A.; Hassanien, A.E.; Soliman, M. End-to-end deep learning model for corn leaf disease classification. *IEEE Access* **2022**, *10*, 31103–31115.
28. Sun, L.; He, J.; Zhang, L. CASF-MNet: Multi-scale network with cross attention mechanism and spatial dimension feature fusion for maize leaf disease detection. *Crop Prot.* **2024**, *180*, 106667.
29. Maginga, T.J.; Masabo, E.; Bakunzibake, P.; Kim, K.S.; Nsenga, J. Using wavelet transform and hybrid CNN–LSTM models on VOC & ultrasound IoT sensor data for non-visual maize disease detection. *Heliyon* **2024**, *10*, e26647.

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