Tomato Leaf Spot Segmentation and Disease Degree Classification Based on the Improved UNet Model

Xueke Xu, Liuai Wu, Zhongrong Zhang, Yulin Shen

Abstract—**In agricultural production, accurate and automated assessment of disease severity is crucial for disease management and yield loss prediction. To meet the precise pesticide application requirements for diseased tomato leaves, this study, in response to the complex background images in the environment, proposed a tomato leaf lesion segmentation model MS-UNet based on the improved UNet architecture. The model incorporates three key innovations: First, the MultiReceptive Field Dilated Convolution Module dynamically adjusts dilation rates to expand receptive fields while maintaining lesion continuity, effectively addressing irregular lesion shapes, and improving segmentation accuracy. Second, the Dynamic Hybrid Attention Module models spatial and channel-wise correlations to suppress background interference and enhance focus on lesion regions. Third, the Weighted Cross-Entropy Loss function implements class-specific weighting to resolve pixel-level class imbalance and optimize model performance. For disease severity assessment, quantitative evaluation is achieved by calculating the ratio of lesion pixels to total leaf pixels. Experimental results demonstrate MS-UNet's superior performance on tomato disease datasets, achieving 92.12% mean pixel accuracy, 88.30% mean intersection over union, and 93.7% F1-score - representing improvements of 3.70%, 3.15%, and 4.46% respectively over the baseline model, with a disease severity classification accuracy of 92.08%. This method enables efficient and accurate lesion segmentation with precise severity grading, providing robust technical support for targeted disease control in tomato cultivation with broad applicability in practical agricultural production.**

Index Terms—**Tomato diseases, Graded detection, Expanded convolution, Attention mechanism, Semantic segmentation**

I. Introduction

TOMATO is a globally cultivated economic crop of significant importance, with its yield and quality directly impacting agricultural production efficiency [1]. Disease infections, however, can result in yield losses of up to 25% to 35% annually. Of these, foliar diseases have emerged as a major obstacle to sustainable tomato production because of their quick spread and subtle early indications [2]. Accurate lesion segmentation and quantitative disease severity assessment are of great significance for developing precise pesticide application strategies and reducing chemical misuse [3]. Traditional disease evaluation methods primarily rely on visual inspection or threshold-based image segmentation techniques, yet these approaches suffer from strong subjectivity and poor environmental adaptability. While traditional image processing techniques are vulnerable to interference from leaf textures and backgrounds under complicated illumination conditions, resulting in significant segmentation accuracy loss, manual tests may show error rates of up to 20% [4].

Recent advancements in deep learning have provided new solutions for intelligent plant disease diagnosis. CNN-based classification models achieve over 90% recognition accuracy for tomato diseases under controlled environments [5]. However, classification models do not satisfy the practical criteria for accurate pesticide administration since they can only provide disease category judgments without knowledge of the spatial distribution of lesions. Pixel-level prediction is how semantic segmentation methods get over this restriction. For crop disease segmentation, Yue et al. [6] proposed an improved SegNet network incorporating Conditional Random Fields for crop disease segmentation, achieving 81.26% precision and 70.91% recall in complex backgrounds. Afzaal et al. [7] created an instance segmentation approach for strawberry illnesses based on Mask RCNN, achieving an average precision of 82.43%. The model demonstrated an 8.31% false identification rate for fruits afflicted with powdery mildew, despite reaching 94.94% accuracy for gray mold using ResNet101. Other shortcomings include a large parameter size and poor test set accuracy. The MC-UNet model, developed by Deng et al. [8], combined a SE module with a Multi-scale Convolution Module and achieved 91.32% accuracy on a bespoke dataset. Despite using SoftPool and SeLU to improve performance, under difficult situations, its segmentation accuracy for dense small lesions dropped noticeably. Furthermore, there was insufficient confirmation of the studies' potential to generalize in field settings because they were mostly dependent on the Plant Village public dataset. By combining a patch-based Transformer encoder with a complementary attention mechanism, Li et al. [9] developed the RSegformer model, which outperformed popular segmentation algorithms and achieved 85.38% mIoU with 14.36M parameters. However, its partial image-based lesion area computation was insufficient for a precise evaluation of the severity of whole-leaf illness. Using the CBAM-FF feature fusion module and the SANet attention mechanism,

Manuscript received May 14, 2025; revised September 8, 2025.

This work was supported in part by the National Natural Science Foundation of China under Grant 51567014 and the Gansu Provincial Science and Technology Plan Project under Grant 2JR5RA797.

Xueke Xu is a postgraduate student at the School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China (e-mail: 2045830511@qq.com).

Liuai Wu is an associate professor at the School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China (corresponding author: e-mail: computer158@163.com).

Zhongrong Zhang is an professor at the School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou, 730070, China (e-mail: gslzzhangzhr@126.com).

Yulin Shen is an associate professor at the Key Laboratory of Advanced Computing of GanSu Province, Lanzhou Jiaotong University, Lanzhou, 730070, China (e-mail: snenyl@cc.gs.cn).

Wu et al. [10] presented an enhanced DeepLabV3+-based approach for maple leaf lesion segmentation that achieved 90.23% mIoU and 94.75% MPA on a custom SLSD dataset, which represents 4.55% and 3.4% improvements over the original DeepLabV3+, respectively. The model kept 25.7M parameters without verifying its generalization performance on additional plant disease datasets, even though it used MobileNetV2 for its lightweight design. For tomato leaf disease picture segmentation, Patil et al. [11] developed an Enhanced Radial Basis Function Neural Network model based on Modified Sunflower Optimization Algorithm for tomato leaf disease image segmentation. Utilizing Gaussian filtering and Contrast Limited Adaptive Histogram Equalization for image preprocessing, the method achieved 98.92% segmentation accuracy, though lacking disease severity grading evaluation. Wang et al. [12] proposed a two-stage model combining DeepLabV3+ with U-Net, where the first stage performed leaf segmentation in complex backgrounds and the second stage accomplished lesion segmentation, ultimately grading severity based on lesion area proportion. However, this approach suffered from error accumulation due to serial two-stage processing [13].

Currently, most disease segmentation methods trained on laboratory single-background datasets exhibit limited generalization capability in actual field environments, with universally declining segmentation accuracy severely restricting practical applications [14]. This study proposes MS-UNet, an enhanced UNet architecture specifically designed for tomato disease images in complex field conditions. The model first incorporates a Multi-Receptive Field Dilated Convolution Module that fuses local details with global contextual information through multi-scale dilated convolution, strengthening feature extraction capability for irregular lesions. Secondly, to effectively overcome background interference, a Dynamic Hybrid Attention Module is introduced, modeling correlations across spatial and channel dimensions to capture intra-class pixel responses and channel dependencies, thereby reducing adverse effects from complex backgrounds and enhancing focus on lesion regions. Thirdly, WCE-Loss is employed to assign category-specific weights, making the model pay more attention to small targets and class imbalance during training for further performance improvement. Comprehensive comparisons with FCN, SegNet, UNet, FPN, PSPNet, and DeepLabV3+ validate MS-UNet's segmentation performance. Finally, based on segmentation results, lesion proportion is calculated and disease severity is graded according to American Phytopathological Society APS standards.

II. RELATED WORK

A. MS - UNet

The UNet semantic segmentation model excels at agricultural crop image segmentation tasks. The Ronneberger team proposed a new design in 2015 that improves small sample learning and precision [15]. The network has a unique symmetric U-shaped structure with three important components: an encoder, a decoder, and skip links that form a deep collaboration mechanism. The encoder path uses cascaded 3x3 convolutional layers and 2x2 max-pooling layers to extract multi-scale features. Each downsampling stage reduces spatial resolution and doubles channel depth, resulting in a high-dimensional feature space with a strong semantic representation. Each downsampling stage halves the spatial resolution of feature maps while doubling the channel depth, progressively constructing a high-dimensional feature space with robust semantic representation. The decoder path gradually restores spatial dimensions through upsampling operations, employing transposed convolutions or interpolation methods for feature map reconstruction. Simultaneously, high-resolution features from corresponding encoder layers are concatenated along the channel dimension via skip connections, effectively fusing local details with global semantic information and mitigating spatial information loss caused by pooling operations. The core innovation lies in the skip connections' cross-layer feature fusion mechanism, which enables complementary enhancement of multi-level features through channel-wise concatenation. The concise and efficient design of UNet provides a crucial paradigm for deep learning applications in pixel-level prediction tasks, making it the foundational framework for this study[16]. To improve segmentation accuracy for tomato leaf diseases, this paper proposes the MS-UNet model based on UNet. First, the Multi-Receptive Field Dilated Convolution Module with varying dilation rates is introduced between the encoder and decoder to reduce detail loss caused by downsampling. To address interference from complex backgrounds, the Dynamic Hybrid Attention Module is incorporated into skip connections, modeling correlations across spatial and channel dimensions to capture intra-class pixel responses and channel dependencies, thereby minimizing adverse effects from complex backgrounds and enhancing focus on lesion regions. Additionally, to counteract pixel class imbalance due to the small proportion of leaves and lesions in images, the WCE-Loss function is adopted to assign category-specific weights during training, prioritizing small targets and imbalanced classes for improved performance. Finally, disease severity is quantitatively assessed by calculating the ratio of lesion pixels to leaf pixels in the segmentation results. The network architecture is illustrated in Figure 1.

B. MRF-DCM

In semantic segmentation tasks, deep convolutional neural networks demonstrate powerful image feature extraction capabilities. However, traditional convolutional networks suffer from detail loss due to pooling layers, while dilated convolutions offer a potential solution[17]. Dilated convolutions expand the receptive field without sacrificing spatial resolution by introducing dilation rates [18]. However, they insert zeros between kernel pixels, creating checkerboard-patterned receptive fields that lose adjacent information and cause gridding artifacts. To address this limitation, this study proposes the Multi-Receptive Field Dilated Convolution Module MRF-DCM, with its architecture illustrated in Figure 2. This module aims to prevent gridding effects while efficiently capturing high-resolution feature maps with preserved spatial information. The MRF-DCM consists of three parallel branches employing dilated convolutions with progressively increasing rates of 1, 2, and 5 following a sawtooth-wave increment strategy. Each branch contains four convolutional layers, each followed by batch normalization and Mish activation functions, enhancing nonlinear representation through

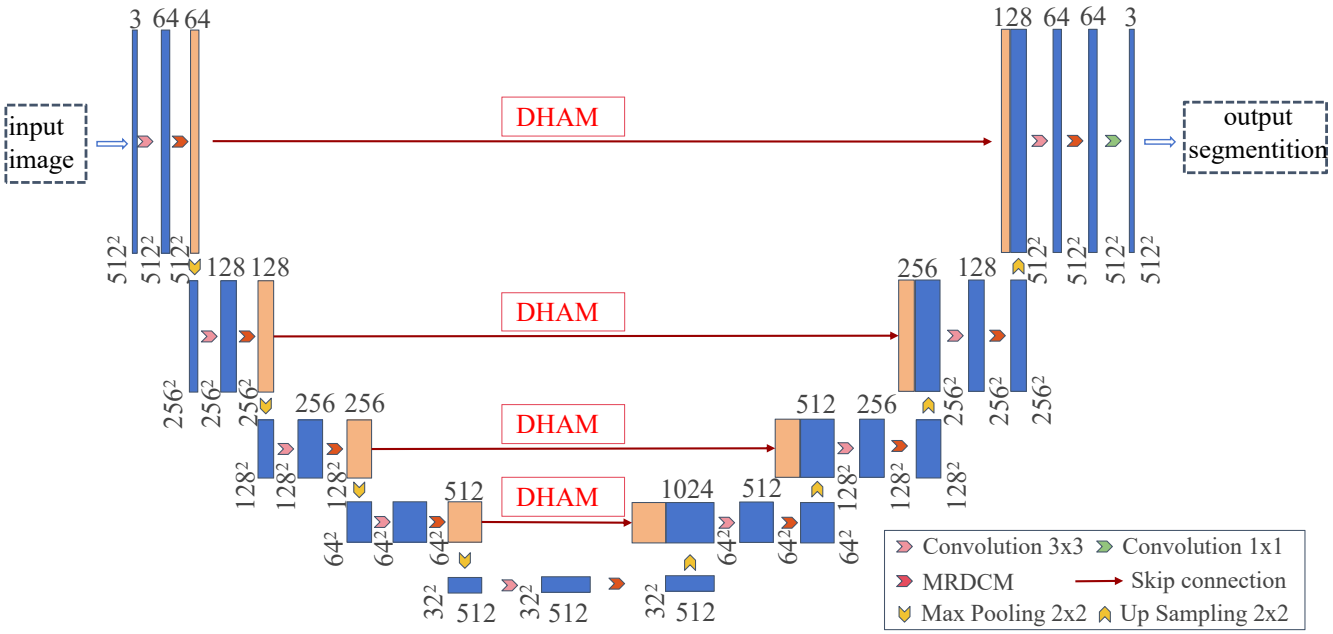

Fig. 1. MS-UNet structure

deep stacking. The design's core advantage lies in multi-scale receptive field synergy: small-dilation convolutions r=1 focus on local details to retain fine-grained features like lesion boundaries; medium-dilation operations r=2 capture regional context by establishing semantic pixel relationships; large-dilation kernels r=5 cover extensive pathological structures to identify widespread diseased areas. Finally, through channel concatenation and 1×1 convolutional fusion, MRFDCM adaptively integrates multi-scale feature responses. This approach simultaneously avoids gridding artifacts while improving the representation of irregularly shaped, multi-size lesions, delivering a more effective solution for semantic segmentation tasks.

C. DHAM

To address the limitations of traditional sequential attention mechanisms in feature interaction and information propagation, this study innovatively proposes a Dynamic Hybrid Attention Module DHAM. Compared to conventional sequential structures like CBAM [19] and ResCBAM which employ channel-spatial dimension cascade processing, DHAM achieves more efficient feature selection capability through multi-scale feature fusion and dynamic gating mechanisms. While traditional methods enhance feature representation through stepwise optimization of different dimensions, their static stacking of attention weights tends to cause over-smoothing of feature responses, particularly leading to significant information attenuation when processing high-frequency details such as lesion boundaries. The proposed DHAM effectively resolves the feature interaction limitations of traditional attention mechanisms via its novel parallel architecture design. As illustrated in Figure 3, DHAM adopts a channel-spatial attention co-optimization mechanism, mainly expressed as follows: First, a dual-path feature aggregation strategy simultaneously utilizes Global Average Pooling and Global Max Pooling to capture global contextual information along the channel dimension [20], where the input feature $F \in \mathbb{R}^{C\times H\times W}$, $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$form a bottleneck-structured MLP with compression ratio r=16, and σ denotes the Sigmoid activation function. As shown in Equation 1:

$$M_c(F) = \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right) \quad (1)$$

Next, a multi-branch parallel structure is constructed after the channel attention module, comprising three branches: 3×3 dilated convolution, 5×5 dilated convolution with dilation rate=2, and 7×7 dilated convolution with dilation rate=3, which respectively capture local detail features and long-range contextual information under different receptive fields. Learnable dynamic weight parameters α, β, and γ are introduced, with Softmax normalization enforcing $\alpha + \beta + \gamma$ =1 to achieve adaptive fusion of multi-scale features, where $F_c = M_c F \otimes F$ represents channel attention-weighted features, and DilConv denotes dilated convolution with dilation rates set to 1, 2, and 3 respectively. As shown in Equation 2:

$$F_{ms} = (\alpha\cdot\text{DilConv}_{3\times3} + \beta\cdot\text{DilConv}_{5\times5} + \gamma\cdot\text{DilConv}_{7\times7})(F_c) \quad (2)$$

Subsequently, a hybrid pooling strategy enhances spatial feature representation, establishing long-range dependencies between spatial positions through 7×7 convolution kernels. As shown in Equation 3:

$$M_s(F_{ms}) = \sigma(\text{Conv}_7[\text{AP}(F_{ms}); \text{MP}(F_{ms})]) \quad (3)$$

Finally, a residual connection structure preserves original feature information to prevent gradient vanishing in deep

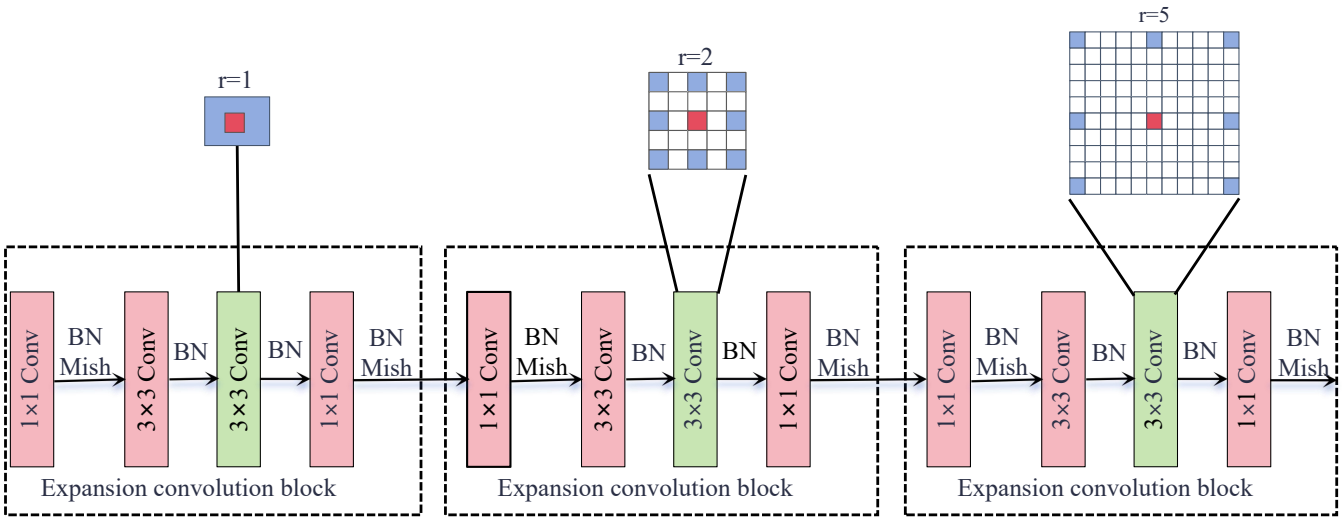

Fig. 2. Illustration of MRF-DCM block

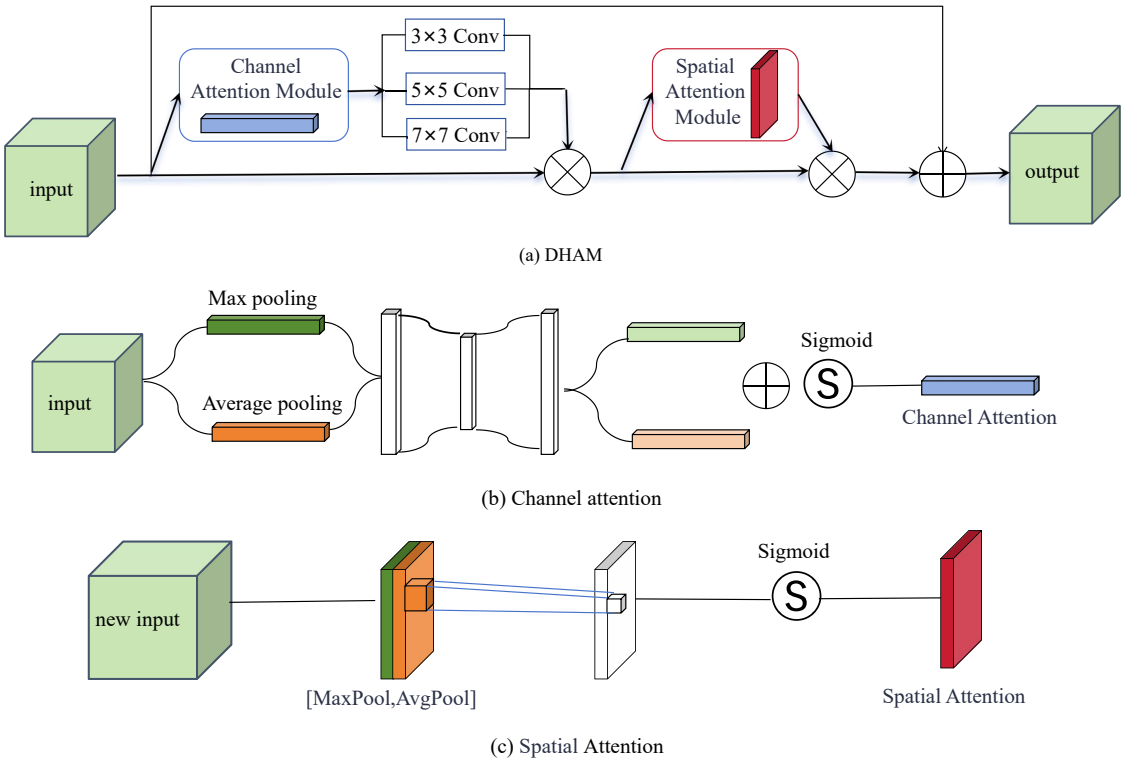

Fig. 3. DHAM attention block

networks. As shown in Equation 4:

$$F_{\text{out}} = F + M_s\left(F_{ms}\right) \otimes F_{ms} \tag{4}$$

This design not only strengthens the model's multi-scale feature extraction capability but also realizes channel-spatial attention co-optimization through dynamic weighting, significantly improving feature discrimination for subtle lesion regions in tomato images under complex scenarios.

D. WCE-LOSS

In the field of agricultural image segmentation, precise segmentation of tomato diseases faces significant challenges of class imbalance [21]. In actual cultivation scenarios, there exists a notable disparity in pixel distribution among background, healthy leaves, and lesion areas. This extreme class imbalance causes traditional segmentation models to predominantly learn features from majority classes while neglecting crucial minority-class lesion features during training. To address this issue, this study innovatively introduces a Weighted Cross-Entropy Loss (WCE-Loss) function, which establishes a pixel-level weighted penalty mechanism based on class importance through the incorporation of class weight coefficients wq [22]. The formula is expressed as:

$$WCE - Loss = -\frac{1}{S}\sum_{i=1}^{S}\sum_{q=1}^{q} w_q y_{iq} \lg p_{iq} \tag{5}$$

The formula is defined as follows where S represents the total number of pixels, q denotes the number of classes in-

cluding background, leaf and lesion, yiq indicates the ground truth label, piq signifies the model's predicted probability for each pixel belonging to the class q, and wq represents the weight for class q. By setting progressive weight ratios of 1:2:3 for "Background", "Leaf", and "Lesion" respectively based on their occurrence frequency and diagnostic importance, this loss function achieves dual optimization objectives: firstly adjusting weights inversely proportional to class frequency by assigning the lowest weight to a high-frequency background, while simultaneously assigning the highest weight to diagnostically critical lesion regions. This weighting strategy enhances the detection rate of small-scale lesions while maintaining overall segmentation accuracy, thereby providing an effective technical solution for precise early disease diagnosis in smart agriculture systems.

III. RESULT AND DISCUSSION

A. Dataset

This study investigates disease segmentation in tomato leaf disease photos. The dataset was collected at the Luoxiang Tomato Production Base in Chaohu City, Hefei, Anhui Province. During data collection, an Apple smartphone was used for natural photography. To ensure image richness and representativeness, photographs were shot from various angles and distances. Each image had to include one entire tomato leaf while maintaining background noise elements like dirt or other leaves. To ensure completeness and diversity, the dataset was collected under diverse weather circumstances, including bright, overcast, and post-rain periods, as well as different times of day. The collection contains photos of two diseases: leaf miner and leaf mold. For subsequent processing and analysis, all images were uniformly resized to 256×256 pixels. After rigorous screening to remove blurry and duplicate images, 680 valid images were ultimately obtained.

LabelImg was employed for pixel-level annotation of the collected tomato disease images to generate corresponding mask maps [23]. During annotation, since images might contain multiple leaves with incomplete leaves in the background, only the central complete leaf was labeled for leaf, and lesion categories. The background, leaves, and two types of lesions, namely Leaf miner and Leaf mold, are respectively labeled as 0, 1, 2, and 3. These manually annotated images serve as the ground truth for measuring segmentation accuracy, with sample annotations shown in Figure 4.

To enhance dataset diversity and improve model robustness, tomato leaf samples were first classified, followed by a series of data augmentation operations including rotation, flipping, local magnification, brightness, and darkness adjustment. Ultimately, the dataset was expanded to 2,150 images. Through random sampling, the dataset was divided into a training set (1,720 images), a validation set (215 images), and a test set (215 images) at an 8:1:1 ratio. Detailed dataset information is presented in Table I.

B. Experimental setup and evaluation metrics

The experimental environment employed the PyTorch framework for model training and testing on a Windows 11 system, with the main platform parameters detailed in Table II.

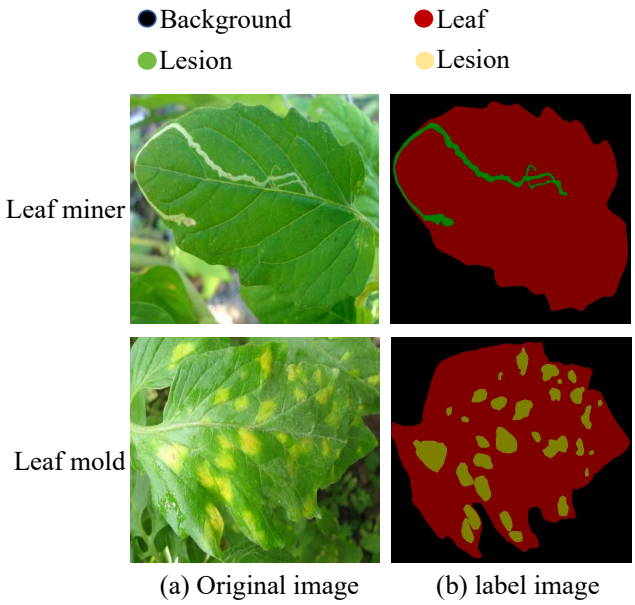

Fig. 4. Image labelling

TABLE I
TOMATO LEAF DISEASE AND PEST DATA SET

Category	Training Set	Validation Set	Test Set	Total
Leaf miner	855	108	106	1,069
Leaf mold	865	107	109	1,081
Total	1,720	215	215	2,150

TABLE II
TRAINING ENVIRONMENT CONFIGURATION

Component	Specification
CPU	AMD Ryzen 9 7940H w/Radeon 780M Graphics
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
Memory	16GB
VRAM	8GB
OS	Windows 11
DL Framework	PyTorch 2.0.1
CUDA	11.7
Python	Python 3.8

TABLE III
TRAINING PARAMETERS CONFIGURATION

Parameter	Value	Parameter	Value
Epochs	100	Optimizer	Adam
Learning rate	0.0001	Weight decay	0.0005
Batch size	8	Momentum	0.937
Image size	512	Workers	4

For model training in this study, all original training set images had a resolution of 256×256 pixels, while the input resolution during training was resized to 512×512 pixels. The training parameter configurations are specified in Table III.

To comprehensively and objectively evaluate the model's segmentation performance, this experiment incorporated five key metrics as primary evaluation criteria: Mean Accuracy,

Mean Intersection over Union, Mean Average Precision, F1-score [24], and Prediction Time, which collectively form the standard for assessing algorithm performance.

1) Mean Accuracy: mAcc quantifies the model's overall classification precision across all categories.

$$mAcc = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} : \quad (6)$$

2) Mean Intersection over Union: mIoU represents the average IoU across all classes, measuring the overlap between predicted and ground truth regions.

$$mIoU = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i + FN_i} \quad (7)$$

3) Mean Average Precision: mAP denotes the mean of AP values for all categories, commonly used in object detection and instance segmentation tasks.

$$mAP = \frac{1}{k}\sum_{t=0.5}^{0.95} AP@t, \quad t \in [0.5 : 0.05 : 0.95] \quad (8)$$

4) F1-score: The F1-score serves as the harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

5) Prediction time/s: The prediction time reflects the real-time processing capability of the model.

$$T_p = \frac{1}{N}\sum_{i=1}^{N} T_i \quad (11)$$

In these metrics, TP, TN, FP, and FN respectively represent true positives, true negatives, false positives, and false negatives; n indicates the number of classes; k denotes the count of IoU thresholds; N represents the number of test samples; and T_i signifies single-sample prediction time.

According to the disease severity classification standard established by the American Phytopathological Society, the severity of disease infection is categorized into five levels based on the percentage of leaf area covered by lesions: Level 0 represents healthy leaves without any lesions; Level 1 indicates sparse patches on the leaf back with coverage area $\leq 5\%$ while maintaining normal leaf morphology; Level 2 corresponds to lesion coverage between 6%-20% accompanied by slight leaf margin curling; Level 3 shows more severe infection with 21%-50% lesion coverage, resulting in leaf deformation and shrinkage; Level 4 represents the most critical condition where lesions cover $> 50\%$ of the leaf area, leading to tissue necrosis or defoliation. This classification system quantitatively evaluates both lesion expansion and tissue damage, providing standardized assessment criteria for disease monitoring, chemical control, and disease-resistant breeding. After annotation, the disease severity index S can be calculated using Equation (12), where P_{lesion} represents the number of lesion pixels in the image; P_{leaf} indicates the count of healthy leaf pixels in the segmentation map; and S stands for the total number of lesion pixels in the image.

$$S = \frac{P_{\text{lesion}}}{P_{\text{lesion}} + P_{\text{leaf}}} \times 100\% \quad (12)$$

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT ATTENTION MODULES

Model	mIoU (%)	mPA (%)	F1-score (%)
UNet	85.15	88.42	89.24
UNet+SE	85.74	89.41	89.64
UNet+ECA	86.03	88.75	89.96
UNet+CBAM	86.14	89.12	90.35
UNet+DHAM	86.45	89.88	90.42

C. Ablation Experiments Result And Discussion

Segmenting disease spots on tomato leaves presents significant challenges due to the complex and variable morphology of lesions, blurred boundaries between healthy and infected tissues, and intricate background interference, which often causes traditional UNet models to overlook subtle lesion regions during feature extraction.

To enhance the model's perception of lesion characteristics, this study employed the UNet architecture as the baseline and compared the improvement effects of various attention mechanisms. Four attention modules were introduced: SE, ECA [25], CBAM, and DHAM, all of which enhanced model performance. As shown in Table IV, the UNet+SE model adopted a channel weighting mechanism to strengthen lesion-related channel features, achieving 0.59 and 0.79 percentage point improvements in mIoU and mPA, respectively, over the baseline. UNet+ECA implemented a cross-channel interaction strategy, showing superior performance to SE with 0.88 and 0.72 percentage point gains in mIoU and F1-score, though its limited spatial detail capture constrained mPA improvement. UNet+CBAM combined dual-path channel and spatial attention, using spatial masks to enhance boundary responses, ultimately reaching 86.14% mIoU and 90.35% F1-score. The proposed UNet+DHAM incorporated a multi-scale dynamic residual attention module that employed parallel multi-scale dilated convolutions to extract both local and global lesion features, dynamically fused spatial responses from different receptive fields and utilized channel attention to filter key feature channels while suppressing background interference. Residual connections maintained deep feature discriminability, enabling precise localization of ambiguous lesions. Experimental results demonstrated mIoU, mPA, and F1-score values of 86.45%, 89.68%, and 90.42%, respectively, representing improvements of 1.30, 1.46, and 1.18 percentage points over the baseline, with optimal comprehensive performance. The DHAM module effectively balanced channel dependency modeling and spatial detail enhancement, offering novel insights for agricultural small-target segmentation. Figure 5 presents the mPA plots of different attention mechanisms.

To validate whether the optimized loss function can enhance model segmentation accuracy, this paper combines UNet with Cross-Entropy Loss (CE-Loss), Focal Loss (F-Loss), Cross Entropy-Dice Mixed Loss (CEM-Loss), and the proposed Weighted Cross-Entropy Loss (WCE-Loss) for performance comparison.

UNet's mPA, mIoU, and F1-score are all lower when using Focal Loss than when using Cross-Entropy Loss, as Table V illustrates. This suggests that although Focal Loss reduces the problems associated with class imbalance, its total seg-

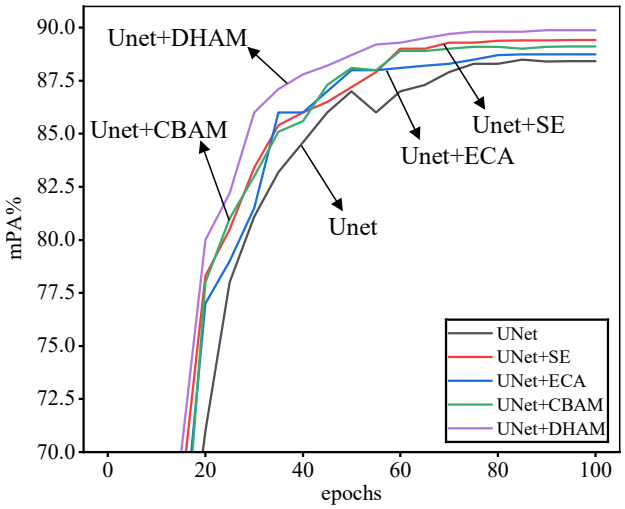

Fig. 5. mPA changes of each attention.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT LOSS FUNCTIONS

Model	mIoU (%)	mPA (%)	F1-score (%)
UNet	85.15	88.42	89.24
UNet+CE-Loss	85.74	88.71	89.39
UNet+F-Loss	85.32	88.68	89.23
UNet+CEM-Loss	86.01	88.76	90.10
UNet+WCE-Loss	86.05	88.99	90.17

mentation performance falls short of the ideal. Compared to utilizing CE-Loss or F-Loss alone, the model shows slight improvements in mPA, mIoU, and F1-score when CEM-Loss is used. The modest increases, however, imply that this hybrid approach has a limited synergistic effect when it comes to simultaneously enhancing region overlap metrics and pixel-wise categorization. Interestingly, using the WCE-Loss allows the model to reach ideal segmentation accuracy. In particular, WCE-Loss improves mPA by 0.28, 0.31, and 0.23 percentage points, respectively, mIoU by 0.31, 0.73, and 0.04 percentage points, and F1-score by 0.78, 0.94, and 0.07 percentage points, respectively, in comparison to CE-Loss, F-Loss, and CEM-Loss. These findings show that WCE-Loss greatly improves the segmentation accuracy of the UNet model for leaf lesions by allowing the model to segment leaf and lesion regions more successfully.

To thoroughly investigate the specific impact of improvement measures on UNet's performance, this study conducted ablation experiments with the following procedure: As shown in Table VI, first, using UNet as the baseline model achieved mIoU, mPA and F1-score of 85.15%, 88.42% and 89.24% respectively; second, when individually incorporating the MRF-DCM into the baseline model, these three metrics improved by 1.57, 0.92 and 1.27 percentage points to reach 86.72%, 89.34% and 90.51% respectively; third, after separately adding the DHAM, all metrics increased by 1.30, 1.46 and 1.18 percentage points compared to the baseline, reaching 86.45%, 89.38% and 90.42%; fourth, when optimizing the loss function with WCE-Loss, the metrics improved by 0.90, 0.57, and 0.93 percentage points to 86.05%, 90.99%, and 90.17%. When implementing module combinations, the joint application of MRF-DCM and DHAM significantly enhanced mIoU, mPA, and F1-score by 2.47, 2.27, and 2.92 percentage points over the baseline, achieving 87.62%, 90.69%, and 92.16%; while the combination of MRF-DCM and WCE-Loss improved these metrics to 87.31%, 90.43% and 91.82%. Ultimately, the complete model (MRDCM + DHAM + WCE-Loss) achieved optimal performance with the three metrics reaching 88.30%, 92.12%, and 93.70%, representing overall improvements of 3.15, 3.70, and 4.46 percentage points over the baseline. Experimental results demonstrate that MRF-DCM effectively captures continuous features of irregular lesions through multi-scale receptive field fusion; DHAM utilizes spatial-channel dual-dimensional attention mechanisms to suppress background interference; while WCE-Loss alleviates class imbalance through dynamic weight allocation. Their synergistic effect significantly enhances lesion segmentation accuracy and model robustness in complex scenarios.

Figure 6 displays performance comparison curves between UNet and MS-UNet during training, showing variations in mIoU and mAP metrics along with training loss reduction. As seen by the data, MS-UNet outperforms baseline UNet on all metrics. In the first training phase, MS-UNet's mAP metric showed a significant improvement in model accuracy. The mAP reached 80% when the training rounds reached 20, and the ultimate stable value was 92.12%, which was 3.7 percentage points greater than the traditional UNet model's figure. In the mIoU statistic, MS-UNet again performed exceptionally well, achieving 88.3%, which was 3.15 percentage points higher than UNet's 85.15%. According to training dynamics, UNet needed more training cycles to reach equivalent results, but MS-UNet's loss function converged to roughly 0.15 by epoch 15. Training dynamics reveal MS-UNet's loss function converged to approximately 0.15 by epoch 15, whereas UNet required longer training cycles to achieve comparable convergence. Notably, MS-UNet demonstrated superior stability in later training stages, with significantly smaller metric fluctuations than UNet.

D. Comparison with other algorithms

This study used several assessment measures, such as mIoU%, mPA%, F1-score, Precision, and Prediction time/s, to assess the segmentation performance of the MS-UNet model for tomato leaf disease spots. The MS-UNet model was compared against the FCN, SegNet, PSPNet, DeepLabV3+, FPN, Swin Transformer, and UNet models under the same conditions [26]. According to Table VII, MS-UNet significantly outperformed conventional models, achieving 88.30%, 92.12%, and 93.70% in the three key measures of mIoU, mPA, and F1-score, respectively. Compared with the FCN model, it showed enhancements of 14.03, 7.99, and 8.33 percentage points in mIoU, mPA, and F1-score respectively; relative to SegNet, the improvements were 9.37, 9.36, and 10.41 percentage points; compared with PSPNet and DeepLabV3+, mIoU increased by 11.91 and 7.07 percentage points, while mPA improved by 10.27 and 6.21 percentage points respectively; even when compared with the classical UNet model, it still achieved gains of 3.15, 3.70, and

TABLE VI
ABLATION EXPERIMENTS RESULTS

UNet	MRF-DCM	DHAM	WCE-Loss	mIoU(%)	mPA(%)	F1-score(%)
✓	–	–	–	85.15	88.42	89.24
✓	✓	–	–	86.72	89.34	90.51
✓	–	✓	–	86.45	89.88	90.42
✓	–	–	✓	86.05	88.99	90.17
✓	✓	✓	–	87.62	90.69	92.16
✓	✓	–	✓	87.31	90.43	91.82
✓	✓	✓	✓	88.30	92.12	93.70

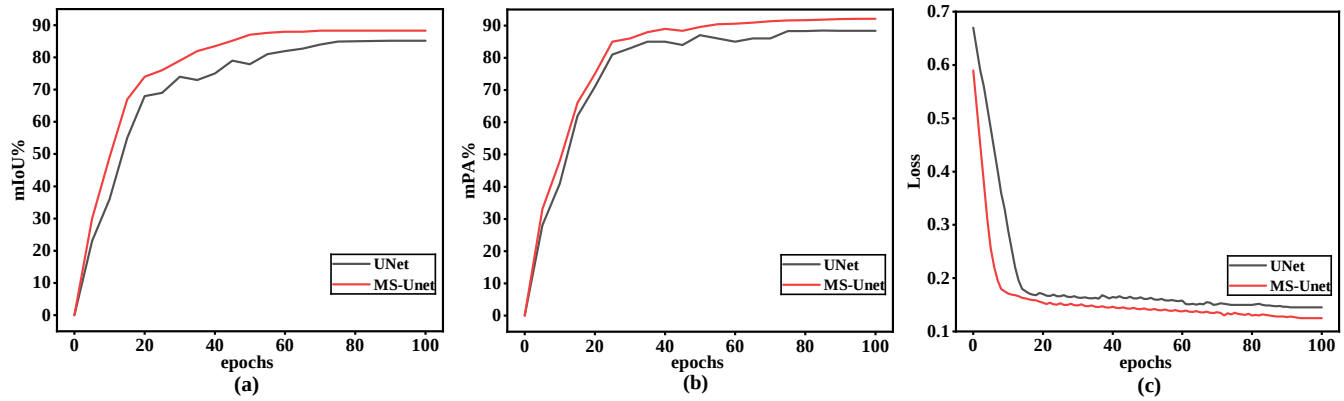

Fig. 6. Changes in the three indicators of each model.

TABLE VII
COMPARATIVE TESTS OF DIFFERENT MODELS

Model	mIoU%)	mPA(%)	F1-score	Precision(%)	Prediction(s)
FCN	74.27	84.13	85.37	86.26	7.89
SegNet	78.93	82.76	83.29	84.68	12.53
PSPNet	76.39	81.85	81.37	83.24	8.21
DeepLabV3+	81.23	85.91	85.46	86.03	8.75
FPN	82.62	86.69	87.23	88.31	10.31
Swin Transformer	86.87	89.83	90.65	92.23	20.13
UNet	85.15	88.42	89.24	89.75	9.01
MS-UNet	88.30	92.12	93.70	94.31	9.23

4.46 percentage points. In terms of segmentation efficiency, MS-UNet's prediction time was 9.23 seconds, showing a slight increase compared to FCN and PSPNet models, but this time cost is justified by the priority of accuracy over real-time requirements in disease spot segmentation tasks.

Figure 7 presents the performance comparison of mIoU versus epoch during training for FCN, SegNet, PSPNet, DeepLabV3+, UNet, and MS-UNet models. MS-UNet demonstrated the most outstanding performance, exhibiting both the fastest mIoU improvement rate and the highest final accuracy, reaching 70 mIoU by epoch 15 and stabilizing around epoch 55, significantly outperforming other models. UNet and DeepLabV3+ showed relatively good but secondary performance, while FCN and SegNet exhibited comparatively slower convergence speeds and lower final accuracy. Overall, MS-UNet demonstrates remarkable advantages in both convergence speed and segmentation accuracy, validating its superiority in handling agricultural image segmentation tasks and providing a more reliable technical solution for intelligent disease diagnosis in agriculture.

E. Visualization results and discussion

Semantic segmentation and feature visualization were carried out on tomato leaves infested with leaf miners and leaf mold to further assess the segmentation performance of FCN,

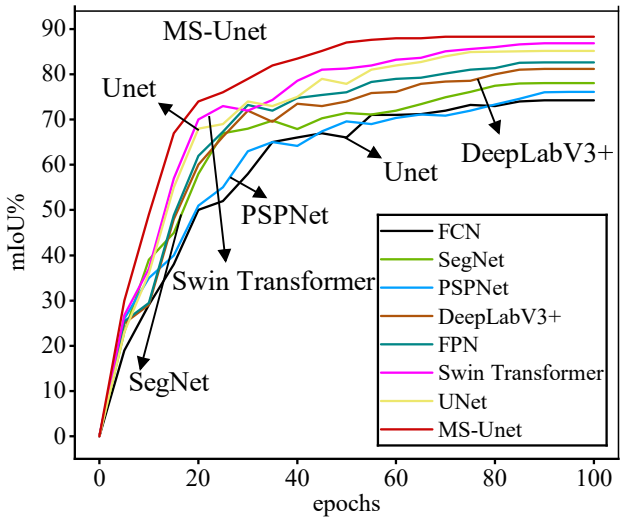

Fig. 7. mIoU changes of each Model.

SegNet, PSPNet, DeepLabV3+, UNet, and MS-UNet. The outcomes are displayed in Figure 8. When the segmentation findings for these two diseases were compared, it was found that the standard FCN model performed poorly at leaf edges and had serious flaws when processing complicated backdrops. Despite outperforming FCN, SegNet, and PSPNet still had some significant drawbacks: While PSPNet demonstrated poor accuracy in lesion area segmentation and blatant oversegmentation issues, SegNet was imprecise in border segmentation under complicated backgrounds. Although UNet and DeepLabV3+ enhanced leaf segmentation performance, they were still limited in their ability to process tiny lesions. The suggested MS-UNet model, on the other hand, showed better segmentation performance, successfully lowering misclassification and under-segmentation errors while preserving outstanding edge segmentation outcomes. By considerably increasing segmentation accuracy for both leaves and lesions, demonstrating greater robustness against background noise, and noticeably lowering the incidence of over-segmentation, this model effectively addressed the problem of inadequate precision in previous models for tomato leaf disease segmentation.

F. Cross-Dataset Algorithm Comparison

To validate the generalization capability of the MS-UNet algorithm, this study conducted comparative experiments against FCN, SegNet, PSPNet, DeepLabV3+, and UNet disease types and their substantial scale, we extracted and organized images of four tomato diseases—Early Blight, Late Blight, Target_Spot, and Yellow Leaf Curl Virus—alongside healthy images. This curated subset was designated as PlantVillage-4 and underwent comprehensive annotation. As shown in Table VIII, MS-UNet achieved the highest IoU across all four disease categories, attaining a mIoU of 85.8%, significantly outperforming other methods ($p < 0.01$). These results demonstrate MS-UNet's superior performance on the PlantVillage-4 dataset, confirming its broad applicability for tomato disease detection.

G. Severity grading of diseases

Significant variations were observed among different deep learning models when grading the severity of tomato leaf mold and leaf miner infections. As shown in Table VII. For healthy leaves (Level 0), PSPNet, DeepLabV3+, UNet, and MS-UNet all achieved 100% accuracy, while SegNet attained a high accuracy of 99.3%. In early-stage infections (Level 1), MS-UNet reached an accuracy of 96.3%, significantly outperforming SegNet. As disease severity increased, performance disparities became more pronounced, For Level 2 infections, MS-UNet achieved 89.7% accuracy, surpassing FCN by 29.3 percentage points. For Level 3, MS-UNet outperformed PSPNet by 13.7 percentage points. In Level 4, MS-UNet attained the highest accuracy. Comparative analysis revealed that MS-UNet achieved an overall average grading accuracy of 92.48% significantly exceeding that of UNet, DeepLabV3+, PSPNet, SegNet, and FCN. These experimental results demonstrate that MS-UNet delivers optimal performance across all severity levels, with particularly superior capability in high-grade disease identification, validating its effectiveness for precise disease severity classification in tomato cultivation.

IV. Conclusion

This study proposes the MS-UNet model for tomato leaf disease segmentation and severity grading, integrating a Multi-Receptive Field Dilated Convolution Module, a Dynamic Hybrid Attention Module, and a Weighted Cross-Entropy Loss to enhance segmentation accuracy and severity assessment. The key findings are summarized as follows:

1) MRF-DCM effectively captures multi-scale contextual features by combining local details and global information through dilated convolution, addressing irregular lesion shapes and scale variations. Experimental results demonstrate that this module improves the model's mIoU and F1-score by 1.57% and 1.27%, respectively, confirming its superior feature representation capability for complex disease spots.

2) DHAM enhances feature discriminability by dynamically adjusting spatial and channel-wise weights, suppressing background interference while highlighting lesion regions. Compared to SE, ECA, and CBAM, DHAM increases the model's mPA by 1.46% and achieves an F1-score of 90.42%. Additionally, WCE-Loss assigns class-specific weights to address small targets and class imbalance, further improving model performance.

3) Compared with models such as FCN and SegNet, MS-UNet performed outstandingly in the task of tomato disease segmentation. The mIoU, mPA, and F1-score reached 88.30%, 92.12%, and 93.70% respectively, which were significantly improved compared with the benchmark UNet. For severity grading, it attains an average accuracy of 92.48%, with a maximum improvement of 64%, demonstrating its effectiveness in precision agriculture applications.

In conclusion, MS-UNet enables high-precision lesion segmentation and severity assessment in complex backgrounds, supporting tomato disease diagnosis and targeted pesticide application. Future work will focus on optimizing environmental adaptability, expanding datasets, improving accuracy, and exploring lightweight deployment.

Declarations:

All authors declare that they have no conflicts of interest.

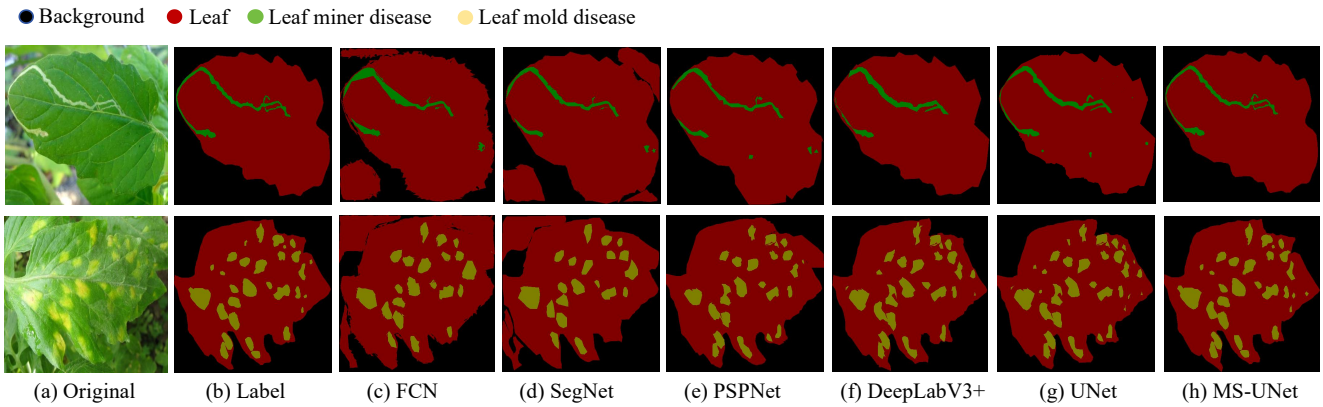

Fig. 8. Comparison of visual results of different models.

TABLE VIII
COMPARISON OF DIFFERENT ALGORITHMS IN PLANTVILLAGE-4 DATASET

Method	FCN mIoU/%	SegNet mIoU/%	PSPNet mIoU/%	DeepLabV3+ mIoU/%	UNet mIoU/%	MS-UNet mIoU/%
All	73.1	75.4	77.2	80.8	84.1	85.8
Early Blight	62.5	74.2	86.5	89.0	88.3	91.2
Late Blight	79.2	70.1	88.3	86.7	82.1	82.8
Target_Spot	76.8	82.4	91.0	88.2	89.5	89.1
Yellow Leaf Curl Virus	84.5	83.2	93.5	95.0	93.8	95.6

TABLE IX
COMPARISON OF CLASSIFICATION ACCURACY RATES OF DIFFERENT MODELS

Disease Level	Images number	FCN Accuracy (%)	SegNet Accuracy (%)	PSPNet Accuracy (%)	DeepLabV3+ Accuracy (%)	UNet Accuracy (%)	MS-UNet Accuracy (%)
Level 0	10	98.1	99.3	100	100	100	100
Level 1	50	83.2	76.5	86.8	90.1	92.6	96.3
Level 2	78	60.4	63.2	73.5	76.4	82.1	89.7
Level 3	83	50.9	56.7	65.2	70.3	72.6	78.9
Level 4	80	80.6	83.7	87.9	89.6	91.7	97.5
Average	–	74.84	76.02	82.68	85.28	87.8	92.48

REFERENCES

[1] F. Gao, H. Li, X. Mu, Y. Zhang, Y. Gao, and Y. Liu, "Effects of organic fertilizer application on tomato yield and quality: A meta-analysis," *Appl. Sci.*, vol. 13, no. 4, p. 2184, 2023.

[2] H. Hong, J. Lin, and F. Huang, "Tomato disease detection and classification by deep learning," in *Proc. Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, 2020, pp. 25–29.

[3] S. Mukhopadhyay, M. Paul, R. Pal, and D. De, "Tea leaf disease detection using multi-objective image segmentation," *Multimedia Tools Appl.*, vol. 80, pp. 753–771, 2021.

[4] D. Kumar and V. Kukreja, "Image segmentation, classification, and recognition methods for wheat diseases: Two decades' systematic literature review," *Comput. Electron. Agric.*, vol. 221, p. 109005, 2024.

[5] M. Agarwal, S. K. Gupta, and K. K. Biswas, "Development of efficient CNN model for tomato crop disease identification," *Sustain. Comput.: Inform. Syst.*, vol. 28, p. 100407, 2020.

[6] Y. Yue, X. Li, H. Zhao, and S. Zhang, "Image segmentation method of crop diseases based on improved SegNet neural network," in *Proc. IEEE Int. Conf. Mechatron. Autom. (ICMA)*, 2020, pp. 1986–1991.

[7] U. Afzaal, B. Bhattarai, Y. R. Pandeya, and J. Lee, "An instance segmentation model for strawberry diseases based on mask R-CNN," *Sensors*, vol. 21, no. 19, p. 6565, 2021.

[8] Y. Deng, H. Xi, G. Zhou, X. Chen, and L. Wang, "An effective image-based tomato leaf disease segmentation method using MC-UNet," *Plant Phenomics*, vol. 5, p. 0049, 2023.

[9] Z. Li, P. Chen, L. Shuai, W. Jiang, and Y. Chen, "A copy paste and semantic segmentation-based approach for the classification and assessment of significant rice diseases," *Plants*, vol. 11, no. 22, p. 3174, 2022.

[10] P. Wu, M. Cai, X. Yi, Y. Li, and Q. Zhang, "Sweetgum Leaf Spot Image Segmentation and Grading Detection Based on an Improved DeeplabV3+ Network," *Forests*, vol. 14, no. 8, p. 1547, 2023.

[11] M. A. Patil and M. Manohar, "Enhanced radial basis function neural network for tomato plant disease leaf image segmentation," *Ecol. Inform.*, vol. 70, p. 101752, 2022.

[12] C. Wang, P. Du, H. Wu, D. Li, and J. Zhao, "A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net," *Comput. Electron. Agric.*, vol. 189, p. 106373, 2021.

[13] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, "Leafgan: An effective data augmentation method for practical plant disease diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1258–1267, 2020.

[14] T. Olivoto, S. M. P. Andrade, and E. M. Del Ponte, "Measuring plant disease severity in r: Introducing and evaluating the pliman package," *Trop. Plant Pathol.*, vol. 47, no. 1, pp. 95–104, 2022.

[15] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, and Q. Tian, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218.

[16] S. Bhagat, M. Kokare, V. Haswani, P. Hambarde, and R. Kamble, "Eff-UNet++: A novel architecture for plant leaf segmentation and counting," *Ecol. Inform.*, vol. 68, p. 101583, 2022.

[17] S. Iqbal, A. N. Qureshi, J. Li, T. Mahmood, and S. S. R. Abidi, "On the analyses of medical images using traditional machine learning techniques and convolutional neural networks," *Arch. Comput. Methods Eng.*, vol. 30, no. 5, pp. 3173–3233, 2023.

[18] R. Gao, "Rethinking dilated convolution for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4675–4684.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[20] F. Bieder, R. Sandkühler, and P. C. Cattin, "Comparison of methods generalizing max-and average-pooling," *arXiv preprint arXiv:2103.01746*, 2021.

[21] M. Shoaib, T. Hussain, B. Shah, R. Ali, and S. W. Baik, "Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease," *Front. Plant Sci.*, vol. 13, p. 1031748, 2022.

[22] Ö. Özdemir and E. B. Sönmez, "Weighted cross-entropy for unbalanced data with application on covid x-ray images," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, 2020, pp. 1–6.

[23] D. Kuznichov, A. Zvirin, Y. Honen, and R. Kimmel, "Data augmentation for leaf segmentation and counting tasks in rosette plants," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 0–0.

[24] J. Ma, Y. Li, H. Liu, Y. Du, and C. Yi, "Improving segmentation accuracy for ears of winter wheat at flowering stage by semantic segmentation," *Comput. Electron. Agric.*, vol. 176, p. 105662, 2020.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

[26] Y. Wang, L. Yang, X. Liu, Z. Chen, and H. Zhang, "An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+," *Sci. Rep.*, vol. 14, no. 1, p. 9716, 2024.