Hybrid Model based on CNN-Transformer for Tomato Pest and Disease Identification

Yu-Cheng Li, Rui Li, and Jia-Bing Zhu

***Abstract*—Convolutional neural networks (CNNs) and Vision Transformer (ViT) fusion models have been widely applied in image recognition. This paper presents an enhanced fusion method based on convolutional and Transformer representation learning (CTRL-F) fusion to improve the model's performance in representing position information and enhance computational efficiency. First, a partial-channel rectangular self-calibration module (PRCM) is designed by combining a rectangular self-calibration module (RCM) with partial convolution (PConv). This module enhances target recognition for objects with varying shapes, reduces computational complexity, and maintains feature extraction capabilities. Second, the Transformer employs a flexible relative position encoding (RPE) instead of fixed positional encoding, enabling the model to capture spatial relationships among objects more accurately across images of varying sizes and scales. Finally, the proposed model achieves an accuracy of 99.83% on the Plant Village tomato pest dataset, which is 3.53% higher than the original method. Comparative results demonstrate that the proposed method outperforms ResNet 50 (98.80%), ResNet 101 (97.14%), and ViT (94.80%). These results demonstrate its effectiveness in identifying tomato pests and diseases.**

***Index Terms*—Convolutional Neural Networks, vision transformer, pest and disease recognition, partial convolution, relative position encoding.**

I. Introduction

TOMATO is China's fourth most widely cultivated vegetable, with an annual output of more than 60 million tons. However, due to the impact of pests and diseases, the problem of crop losses has received more attention. According to the Food and Agriculture Organization of the United Nations [1], the global economic losses of the major food and cash crops caused by pests are more than 20%. Hence, pest monitoring and yield forecasting are crucial to ensuring food security. Accurate identification of pests and diseases is essential for implementing effective control measures.

Traditional recognition methods relying on expert experience suffer from low efficiency and poor scalability, making them unsuitable for real-time monitoring in modern agricultural practices. With the rapid advancement of artificial intelligence and computer vision [2], numerous researchers have invested in various methods for pest identification. These methods can generally be classified into conventional machine learning and deep learning. However, traditional machine learning methods perform well in controlled environments but have low accuracy in complex field conditions. Although CNNs perform well in conventional target recognition, their fixed square receptive fields limit their ability to perceive strip-shaped, highly directional, and structurally variable targets (e.g., lesions, insect bites, and corrosion). This limitation makes it challenging to capture spatial features fully. Although ViT achieves superior performance in capturing global dependencies, its limited capacity for local feature extraction makes it prone to overfitting when trained on small-scale datasets.

Using fixed absolute positional embeddings in hybrid models constrains positional encoding, limiting adaptability to complex agricultural scenes with varying image sizes and aspect ratios. To solve these limitations, this paper presents an improved CNN-Transformer fusion model to enhance recognition, improve computational efficiency, and intensify the model's ability to represent positional information.

The main contributions in this paper are concluded as follows: (1) A PRCM based on partial channel convolution is designed to combine an RCM with PConv, which not only improves the recognition ability of different shapes and sizes of targets but also reduces the calculation complexity and maintains a robust feature extraction ability. (2) This paper also replaces the fixed position information in the original Transformer with a more flexible RPE, so that the proposed method can understand the position relationship of the object more accurately when the input images have different sizes and scales. (3) Results indicate that the proposed model can achieve an accuracy of 99.83% on the public dataset of Plant Village [3], which can prevail over the baseline model by 3.53%.

The rest of this paper is organized as follows. In Section II, the related works are briefly reviewed. In Section III, an improved model is proposed to enhance the recognition ability and computation efficiency. Experiments and analyses are provided in Section IV. The conclusion of this paper is completed in Section V.

II. Related Works

Following the rapid development of deep learning, CNNs have successfully changed the way of pest identification. Especially, Praveen and Jung [4] proposed an improved YOLO model combining CBAM attention mechanism with

Manuscript received July 20, 2025; revised October 22, 2025.

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2008085QF329; the Natural Science Foundation for the Higher Education Institutions of Anhui Province under Grant 2023AH051867; the Major Science and Technology Projects of Anhui Provincial Science and Technology Innovation Platform under Grant S202305a12020025; and the key Discipline Construction Project of Anhui Science and Technology University under Grant XKXJGY002.

Yu-Cheng Li is a master's student of the College of Information and Network Engineering, Anhui Science and Technology University, Bengbu 233030, China (e-mail: 17658159475@163.com).

Rui Li is an associate professor of the College of Information and Network Engineering, Anhui Science and Technology University, Bengbu 233030, China (corresponding author to provide phone: 86-17555271029; e-mail: lir@ahstu.edu.cn).

Jia-Bing Zhu is a professor of the School of Electronic Science and Engineering, Huainan Normal University, Huainan 232038, China (e-mail: zjb3617@163.com).

spatial transform network (STN) and thin plate spline (TPS) modules, which can significantly enhance the perception of irregular and distorted objects through spatially adaptive transformation. This method performs superiorly on plant growth phenotype datasets containing complex backgrounds and occlusions. On the other hand, Ung [5] has proposed a multiple CNN fusion method, which can achieve efficient recognition of insect pest classification by combining multiple CNN models for feature fusion [6]. However, although these techniques improve recognition accuracy, they still have key limitations. The receptive field is a fixed square window for local information extraction in the traditional convolution structure. Although this structure performs well in regular target recognition, it permits deficiencies in perception ability for targets with spatial directionality and structural variation, such as strip-shaped disease spots, insect bite tracks, and linear corrosion [7], [8].

To overcome these limitations, the Vision Transformer (ViT) has been proposed as an emerging architecture [9]. ViT addresses the long-range dependency problem using a global self-attention mechanism on image patches and surpasses CNNs on public datasets [10]. Although ViT outperforms CNNs, it still faces two significant challenges in plant pest identification: (1) it needs large training samples to avoid overfitting[11]; (2) its computational complexity increases as image resolution increases. Although ViT exhibits excellent global modeling capability, its local feature extraction remains weak, which can easily lead to overfitting when the sample size is small.

In recent years, many researchers have tried integrating CNN and Transformer structures to balance local feature extraction and global modeling abilities [12]. TransUNet proposed by Chen [13]uses CNN as an encoder to extract fine-grained features, then feeds these features into the ViT to model globally. Although TransUNet achieves good results in medical image segmentation and related tasks, its positional encoding is still limited by fixed absolute embeddings. It is unsuitable for complex agricultural scenes with varying input image sizes and aspect ratios [14].

III. Imporoved CNN-Transformer Hybrid Models using Dynamic Fusion Method

CTRL-F model [15] is a typical CNN and Transformer fusion image classification framework, aiming to balance the local modeling advantages of CNN and the global modeling capabilities. The convolution branching part uses the lightweight mobile inverted bottleneck convolution (MBConv) module as the basic construction unit. This module has good expression ability and structural efficiency, and has the advantage of the Transformer model. In addition, the Squeeze-and-Excitation (SE) attention mechanism has been integrated into MBConv to dynamically recalibrate the feature response along the channel dimension, thereby improving the model's ability to perceive global semantic information.

Although the CTRL-F model enables effective feature extraction and global modeling capabilities in the fusion of CNN and Transformer structures, it still has three main limitations: (1) A limited local receptive field leads to insufficient spatial modeling accuracy [16]. (2) The computational resource consumption in the feature fusion stage is very high [17]. (3) Absolute position coding cannot model relative spatial relationships [18]. To overcome these problems, based on the CTRL-F model, this paper proposes a lightweight and efficient improved CNN-Transformer fusion model displayed in Fig. 1.

In Fig. 1, the yellow blocks represent key improvements proposed in this paper. The input image first passes through an initial convolution stage (S0) containing stem blocks, followed by four convolution stages (S1 to S4), each based on the MBConv structure. Two sets of features are extracted in S2 and S4, respectively, and put into the Transformer module for global modeling [15]. The feature fusion module combines the local structure information captured in CNN branches with the worldwide information obtained by Transformer branches, enhancing the ability to recognize different-scale and shape targets in pest images.

A. RCM with PConv

In traditional convolutional structures, the receptive field is limited to a fixed square window for local information extraction. Although this structure performs well in regular target recognition, it has severe limitations in recognizing pest targets with pronounced spatial directionality and structural variability (such as strip-shaped disease spots, insect bite tracks, and linear corrosion). This limitation is further exacerbated in pest images, where object distributions often exhibit strong geometric anisotropy (such as lateral extension or longitudinal growth); it is difficult for the convolution kernel with a single scale and direction to accurately capture their key features[19].

To enhance the modeling capability of sensitive directional features, a rectangular self-calibration module [20] is employed in the convolution branch. The RCM integrates Rectangular Self-Calibration Attention (RCA), batch normalization (BN), and a multilayer perceptron (MLP) to enhance feature representation [21]. The adaptive receptive field adjustment in spatial direction is achieved by extracting channel importance weights in horizontal and vertical directions, thereby enhancing the model's perception capability of target structural characteristics and edge information [22]. RCM in Fig. 2 first extracts global context information in horizontal and vertical directions through horizontal and vertical pooling, respectively. The two directional features are combined via broadcast addition to construct a rectangular region of interest (ROI). Its mathematical expression is as follows

$$\boldsymbol{P} = H_P(x) \oplus V_P(x), \tag{1}$$

where $\oplus$ denotes broadcast addition, and $\boldsymbol{P}$ is the resulting rectangular area of interest. To better align the ROI with object structures, the RCM introduces a shape self-calibration function that adjusts the horizontal and vertical shapes through two independent band convolutions. Specifically, the horizontal band convolution refines the ROI in the horizontal direction, while the vertical band convolution modifies its vertical counterpart [23]. Finally, the calibrated attention map is generated through the Sigmoid function [20]defined in (2).

$$\xi_C(\bar{y}) = \delta\left(\psi_{k\times 1}\left(\phi\left(\psi_{1\times k}(\bar{y})\right)\right)\right), \tag{2}$$

where $\psi_{k\times 1}$ and $\psi_{1\times k}$ represent horizontal and vertical banded convolutions, ϕ denotes batch normalization and

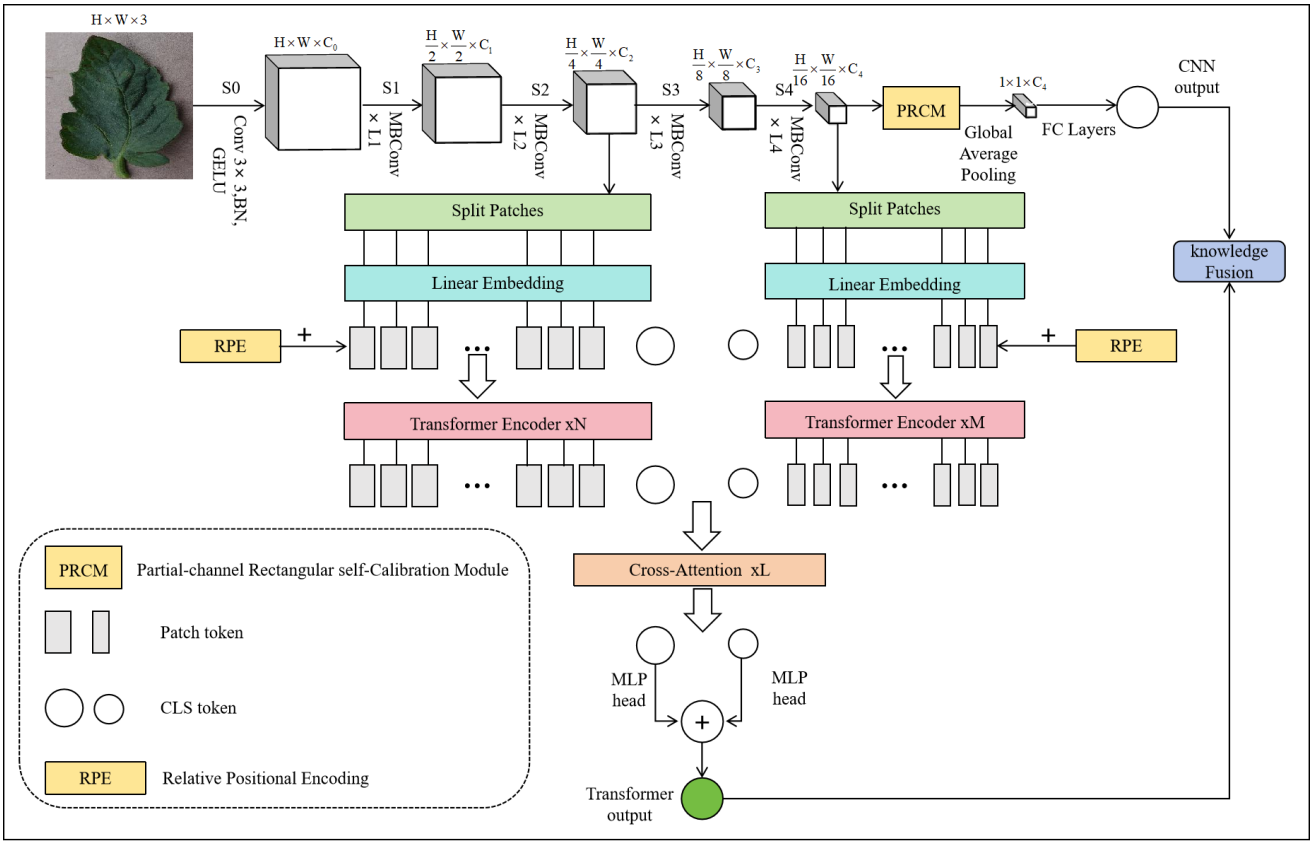


Fig. 1. Improved CNN-Transformer fusion network structure

ReLU activation [24], and δ represents the Sigmoid function. After completing shape self-calibration, RCM has extracted local details from input features using a 3×3 depth-separable convolution and fuses calibrated attention, which can be expressed as

$$\xi_F(x,y) = \psi_{3\times3}(x) \odot y, \tag{3}$$

where $\psi_{3\times3}$ indicates the 3×3 depthwise separable convolution, y corresponds to the calibrated attention feature, and $\odot$ signifies the element-wise (Hadamard) multiplication. Finally, RCM can optimize feature representation using BN and MLP and utilize a residual connection to enhance the feature reuse ability. The output expression [25] can be expressed as follows

$$\boldsymbol{F}_{\text{out}} = \rho\left(\xi_F\left(\boldsymbol{x}, \xi_{\text{C}}\left(\text{H}_{\text{P}}(\boldsymbol{x}) \oplus V_P(\boldsymbol{x})\right)\right)\right) + \boldsymbol{x}, \tag{4}$$

where ρ corresponds to the BN and MLP operations and $\boldsymbol{x}$ denotes the input feature map.

To reduce the computational complexity of the module, this paper introduces a partial channel convolution mechanism in RCA [26]. Based on the redundancy assumption of the feature graph in the channel dimension, this mechanism can effectively reduce computational cost and enhance the model's expressive ability by performing a convolution operation on a subset of the channels. In Fig. 3, the partial channel convolution convolves a part of the channels, and reduces computational complexity while enhancing the model's selectivity for key features [27]. Compared with full-channel convolution, PConv decreases computational overhead and focuses more on capturing representative channel information, thus effectively avoiding redundant computation [28].

Let the denotes feature input $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$, PConv initially splits the channels into two groups: the first $C_p = r \cdot C$ ($r = \frac{1}{4}$) channels are processed by the $K \times K$ convolutional kernel, and the rest $C - C_p$ channels are passed through without any changes. Finally, the outputs of both channel subsets are merged across the channel axis and yield the resultant feature map $\boldsymbol{Y} \in \mathbb{R}^{C\times H\times W}$. The channel selection strategy adopts a fixed sampling method to realize continuous and regular memory access. This method selects several consecutive channels in the input channels as representations to perform a standard convolution operation, while the rest remain unchanged. This selection method simplifies the implementation difficulty and facilitates the optimization of memory access by hardware accelerators. The experiment assumes the input and output feature maps have the same channels. The proportion of partial channels can be expressed as $r = \frac{c_p}{c}$, where c_p denotes how many channels participate in the convolution operation, and c represents the complete channel count. Hence, one can express the floating-point operation of a PConv module [26]as

$$h \times w \times k^2 \times c_p^2, \tag{5}$$

where h and w represent the height and width of the feature map, respectively, and k is the convolution kernel size. For $r = \frac{1}{4}$, the floating-point operations of PConv are reduced to $\frac{1}{16}$ compared to a regular convolution. Additionally, PConv efficiently decreases memory bandwidth requirements and computational cost [26]can be expressed as

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p. \tag{6}$$

For $r = \frac{1}{4}$, its memory access is merely $\frac{1}{4}$ relative to

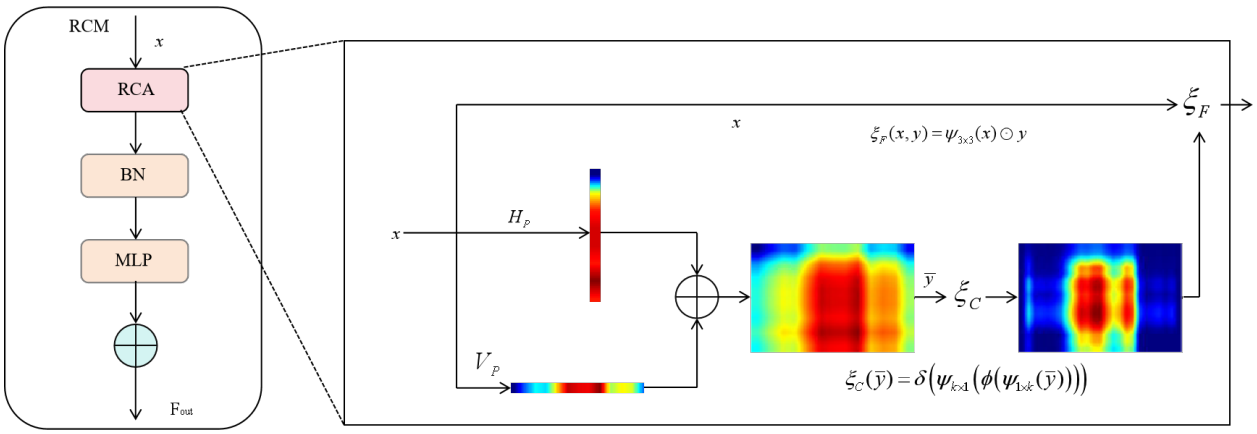


Fig. 2. Architecture of the RCM Module

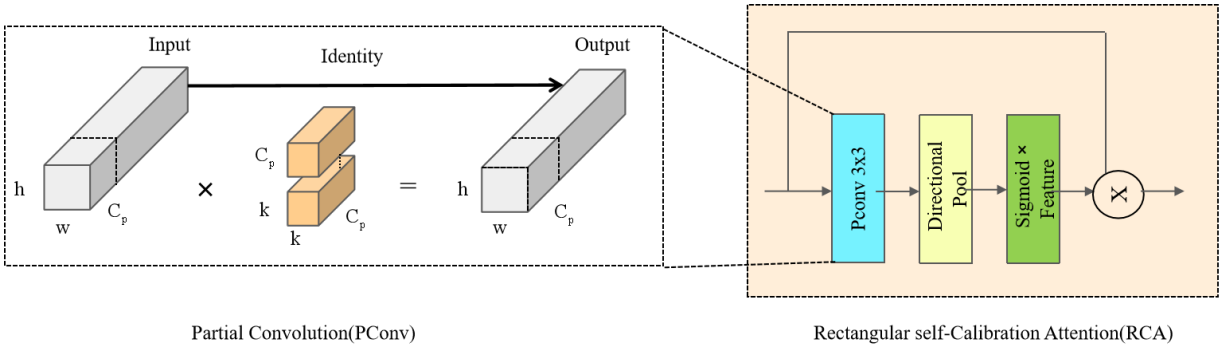


Fig. 3. Architecture of the PConv Module

a standard convolution. PConv has an advantage regarding information retention, as it does not participate in convolution and retains the original global feature information. This processing can help alleviate the feature degradation caused by the overfitting problem. By using point-wise convolution (1×1 PWConv) after PConv, the model can fuse information from all channels to enhance its overall expressive ability.

B. Incorporating relative position encoding into vision transformers

To improve the sensitivity of the Transformer model to positional relationships in sequences, we introduce relative position-coding (RPE) into the cross-attention mechanism [29]. Relative position coding strengthens the Transformer's capacity to model the relative relationships between elements in a sequence. Unlike traditional absolute position coding, it focuses on the "distance" between two tokens rather than the "position" of each in the sequence. By using a learnable bias matrix, the model not only considers the matching degree of Query and Key but also adds a bias term based on their relative positions when calculating the attention score, so that the attention mechanism can perceive and utilize the positional differences between tokens [30]. In addition, the relative position between each pair of tokens is mapped to an offset index, which retrieves the corresponding offset value from the offset matrix and adds it to the attention score. This method ensures the network extracts practical structural features and handles variable-length inputs, enhancing its generalization and positional awareness.

For a sequence of length L, we use a trainable relative position bias matrix $\mathbf{B}\in\mathbb{R}^{(2L-1)\times H}$, with H representing the total number of attention heads. Each row encodes the bias associated with a specific relative position in this matrix. For arbitrary positions i and j ,we define the bias index as

$$\text{index}(i,j)=i-j+(L-1). \tag{7}$$

This operator ensures that the index is non-negative and that all possible relative offsets are fully covered. In addition to the regular dot-product score, we add a relative position bias $\mathbf{b}_{ij}$ to the score matrix so that the final attention distribution reflects the relative relationship between tokens. In the cross-attention module, where $\mathbf{Q}\in\mathbb{R}^{B\times H\times1\times d}$, $\mathbf{K}\in\mathbb{R}^{B\times H\times L\times d}$ and $\mathbf{V}\in\mathbb{R}^{B\times H\times L\times d}$ represent the query, key, and value (B denotes the batch size and d represents the dimension per head), the standard attention score is defined by

$$\mathbf{A}_h(i,j)=\frac{\mathbf{Q}_h(i)\cdot\mathbf{K}_h(j)^\top}{\sqrt{d}}. \tag{8}$$

With the introduction of relative position bias, the attention score was updated to

$$\mathbf{A}_h(i,j)=\frac{\mathbf{Q}_h(i)\cdot\mathbf{K}_h(j)^\top}{\sqrt{d}}+\mathbf{B}_h[\text{index}(i,j)], \tag{9}$$

where $\mathbf{B}_h[\cdot]$ denotes the bias vector for the h head, looked up via the relative position index. We then perform a softmax over each head's score matrix to normalize the attention weights and apply these weights to $\mathbf{V}$ to compute the final output.

IV. Experiments and Analysis

A. Dataset and preprocessing

This paper uses the Plant Village dataset [3], which covers 10 tomato leaf diseases and healthy samples, including bacterial spot, early blight, late blight, leaf mold, wilt spot, two-spotted spider mite, target spot, mosaic virus, chlorotic leaf curl, and healthy leaves [31]. All images present intact tomato leaves against a monochromatic background, reducing noise and occlusion and improving data quality. The resolution of all images is 256×256 pixels. To minimize unnecessary edge information, cropping the image from the center is adopted to resize it to 224×224 pixels. The dataset is divided into three subsets: a training set, a validation set, and a test set, with a 7:2:1 ratio. A typical example image of tomato disease from the Plant Village dataset [3] is shown in Fig. 4.

B. Model evaluation metrics

Accuracy, Precision, Recall, and F_1 were used as the primary evaluation indices to validate the efficacy of the proposed approach, which are defined by (10)-(13) as follows [32].

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

$$\text{Precision } = \frac{TP}{TP + FP}, \tag{11}$$

$$\text{Recall } = \frac{TP}{TP + FN}, \tag{12}$$

and

$$F_1 = \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{precision } + \text{ Recall}}, \tag{13}$$

where TP denotes a true positives, FN for false negatives, FP for false positives, and TN for true negatives [32]. A confusion matrix is an intuitive evaluation tool that shows how well a model predicts various types of samples. This paper uses the matrix form to visualize the results of disease prediction for different disease categories. This method can identify misclassification and provide data to improve the reliability of a model.

C. Ablation experiment

To verify the contribution of each module (RCM, PConv, and RPE) proposed in this paper, ablation experiments were designed to evaluate the model's performance. Based on the original CTRL-F model as a baseline, four comparative experimental models are constructed using a rectangular self-calibration module with partial channel convolution and utilizing relative position coding in the Transformer branch, respectively. The influence of each improvement on the overall recognition performance is then analyzed. Experiments were conducted on the Plant Village tomato pest image dataset, with all experiments using Accuracy, Precision, Recall, and F_1 value. The performance of each model is presented in Table I below. Table II presents a computation comparison of the baseline model and its three variants in terms of parameter count, M-Adds, and FLOPs, from which the impact of each module on the overall computational cost can be directly observed.

The experimental findings, as detailed in Table I and Table II: in the Baseline model, the Accuracy, Precision, Recall, and F1 values are 96.30%, 96.62%, 96.48%, and 96.45%, respectively. Meanwhile, the parameters, M-Adds and FLOPs are 21.48M, 3582.34M, and 7164.68M. After introducing the RCM module, all indexes are significantly improved, and the accuracy rate is enhanced to 98.37%, indicating that the enhancement of direction perception ability substantially impacts the extraction of spatial structure features. Correspondingly, the computational overhead has also increased, with the number of parameters rising to 21.92M, reflecting the additional self-calibration operations introduced by RCM.

After replacing the full channel convolution with partial channel convolution, the model's accuracy improves to 99.41%, and all indices are highly consistent. Meanwhile, the number of parameters related to calculation complexity has been reduced to 21.34 M. This result indicates that partial channel convolution can improve the recognition performance of the model while effectively reducing its computational complexity, thereby verifying its design advantages of both being lightweight and expressive. Based on the above structure, relative position coding is introduced to replace the original absolute position coding in the Transformer [33]. The model's accuracy is improved to 99.83%, and the Precision, Recall, and F1 values are 99.86%, 99.84%, 99.85%. The corresponding computational overhead remains almost unchanged, with the number of parameters being 21.26M. This demonstrates that relative position coding can capture objects' relative spatial structure information in complex scenes.

From Table I, the three improved strategies proposed in this paper significantly enhance spatial perception ability, reduce computational complexity, optimize spatial modeling effects, and exhibit good complementarity and superposition effects on the model structure. The proposed model improved accuracy by 3.53% and reduced FLOPs by 4.97% compared with the baseline on the Plant Village tomato pest dataset, verifying the effectiveness and practical value of the research method.

D. Comparison experiments

To demonstrate the validity of the presented approach, comparative experiments are executed to evaluate its performance against existing methods in the tomato pest recognition task. ResNet-50 [34], ResNet-101 [34], ViT-B/16 [35], ConvNeXt-T [36], EfficientNetV2-S [37], GoogleNet [38], and Swin Transformer [39] are selected to conduct comparative experiments on the Plant Village tomato disease dataset. All models are trained under the same training rounds, learning rate scheduling, and data enhancement strategies to guarantee experimental fairness. Table III demonstrates the performance of different models in the tomato pest identification task, including Accuracy, Precision, Recall, and F1 Score. Table IV compares the computational complexity of several representative models with the proposed method regarding parameter count, M-Adds, and FLOPs.

In Table III, the proposed model achieves the highest value in all evaluation indexes, with an accuracy of 99.83%,

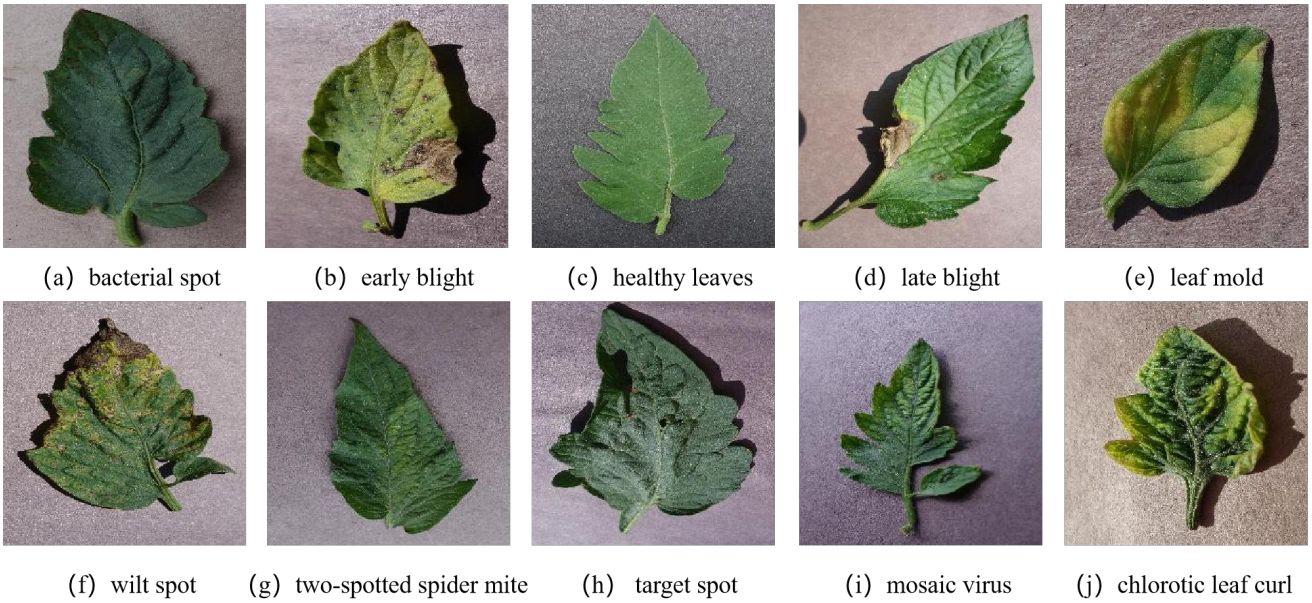


Fig. 4. Examples of Tomato Pest and Disease Images

TABLE I
ABLATION EXPERIMENTS

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Baseline	96.30	96.62	96.48	96.45
Baseline + RCM	98.37	98.47	98.45	98.44
Baseline + PRCM	99.41	99.45	99.43	99.43
Baseline + PRCM + RPE	**99.83**	**99.86**	**99.84**	**99.85**

TABLE II
COMPUTATION COMPARISON OF BASELINE AND VARIANTS

Models	Params (M)	M-Adds (M)	FLOPs (M)
Baseline	21.48	3582.34	7164.68
Baseline + RCM	21.92	3810.78	7621.56
Baseline + PRCM	21.34	3375.11	6750.22
Baseline + PRCM + RPE	**21.26**	**3404.45**	**6808.91**

TABLE III
EXPERIMENTAL RESULTS OF RECOGNITION FOR DIFFERENT NETWORK METHODS

Models	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)
ResNet-50 [34]	98.80	98.88	98.85	98.86
ResNet-101 [34]	97.14	97.30	97.28	97.23
ViT-B/16 [35]	94.80	95.00	94.80	94.70
ConvNeXt-T [36]	98.80	98.80	98.80	98.80
EfficientNetV2-S [37]	98.07	98.18	98.16	98.15
GoogleNet [38]	97.04	97.14	97.18	97.07
Swin Transformer [39]	95.85	96.39	96.05	96.07
Proposed method	**99.83**	**99.86**	**99.84**	**99.85**

which is 1.03% higher than that of the best contrast models (ResNet-50 and ConvNeXt-T). Meanwhile, the precision, recall, and F_1-score are 99.86%, 99.84%, 99.85%, indicating the model has high stability and generalization ability in the disease classification task. Compared with CNNs, the proposed model improves accuracy by 1.03% over the best contrast CNNs (ResNet-50 and ConvNeXt-T) and by 2.79% over GoogleNet. Compared with Transformers, the proposed model improves accuracy by 5.03% over ViT-B/16 and 3.98% over Swin Transformer, indicating that our method has better adaptability on small-scale datasets.

From the computational efficiency perspective, as shown in Table IV, the proposed model achieves a favorable trade-off between computation and model size. It has FLOPs of 6808.91M and parameter size 21.26M, which are the lowest (or among the lowest) values for the high-performing models. GoogleNet has the smallest footprint overall (10.33M params, 3194.38M FLOPs), but its recognition performance

TABLE IV
COMPUTATION COMPARISON OF DIFFERENT NETWORK METHODS

Models	Params (M)	M-Adds (M)	FLOPs (M)
ResNet-50 [34]	23.53	4131.72	8263.43
ResNet-101 [34]	42.52	7864.41	15728.82
ViT-B/16 [35]	85.65	16862.87	33725.74
ConvNeXt-T [36]	27.81	4454.77	8909.55
EfficientNetV2-S [37]	20.19	5397.22	10794.43
GoogleNet [38]	**10.33**	**1597.19**	**3194.38**
Swin Transformer [39]	27.53	4380.27	8760.53
Proposed method	21.26	3404.45	6808.91

is substantially lower (Accuracy 97.04%). EfficientNetV2-S has slightly fewer parameters than ours but notably higher FLOPs. Under low computational overhead, the model still obtains recognition performance, demonstrating a good balance between accuracy and efficiency, offering a more reliable solution for disease detection.

From the training curve in Fig. 5, the model rapidly improves verification accuracy within the first 10 to 20 epochs, indicating that it can efficiently capture the discriminant information of the data at an early stage. During 20 and 60 epochs, the accuracy exhibits a steady upward trend and indicates the proposed model achieves substantial generalization on the training samples, effectively avoiding the overfitting problem. As training progressed, the model continued to optimize, eventually reaching or slightly exceeding the level of other advanced models, such as ConvNeXt-T, EfficientNetV2-S, the ResNet family, ViT-B/16, GoogleNet, and Swin Transformer. The results demonstrate that the fusion model responds rapidly in the initial stage and maintains a stable and effective learning process throughout the training process, providing a powerful solution for related tasks with high learning speed and excellent final performance.

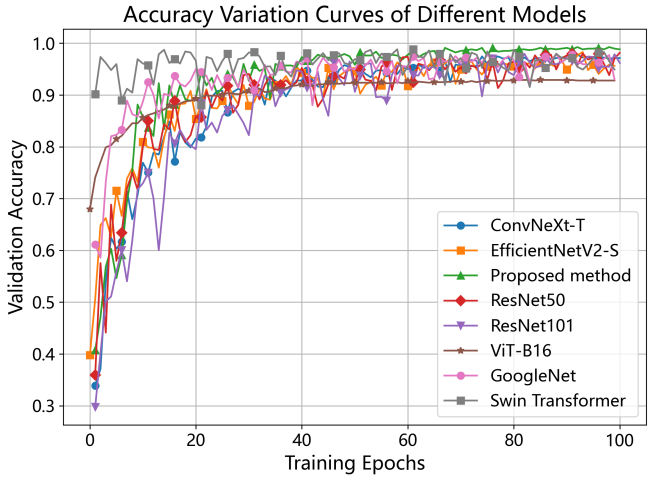


Fig. 5. Accuracy variation curves for different models

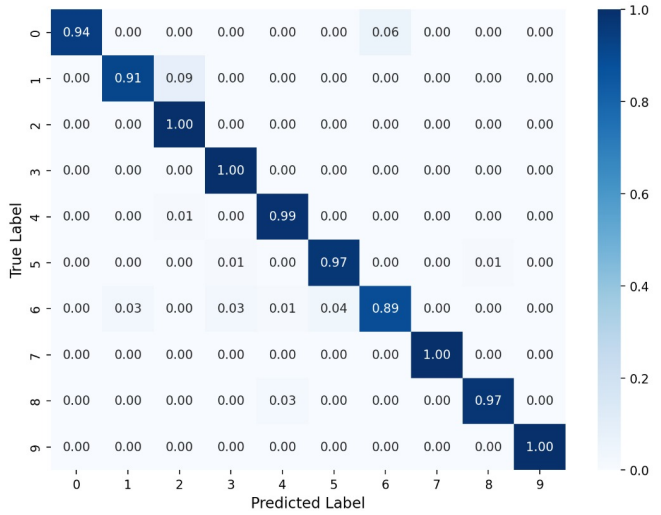
The confusion matrix is a key metric for evaluating the model's classification accuracy. Within the confusion matrix, actual classes are represented by rows and predicted classes by columns. In Figs. 6, 7, it is evident that although ResNet101, ResNet50, EfficientNetV2-S, ViT-B/16, ConvNeXt-T, GoogleNet, and Swin Transformer models all exhibit distinct classification abilities, the proposed model in this paper stands out. Its confusion matrix shows that the prediction accuracy of all categories is close to perfect, and there is no misjudgment phenomenon. This demonstrates that the proposed model efficiently extracts fine-grained features and captures subtle differences between categories, particularly in distinguishing disease categories with similar characteristics.

V. CONCLUSION

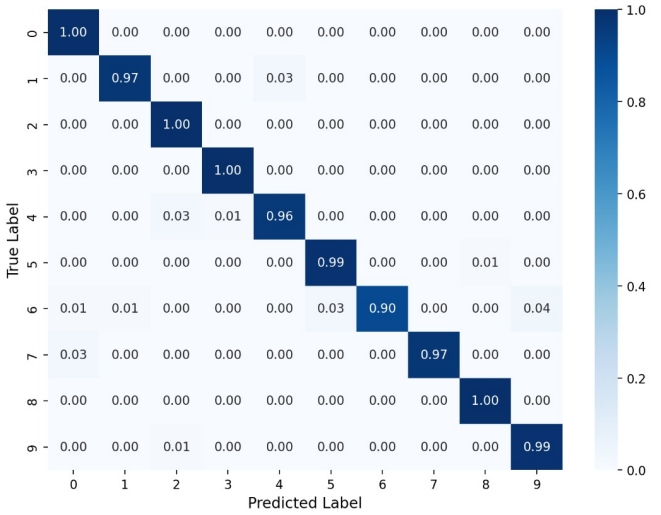
An improved CNN-Transformer fusion model with high performance is proposed based on the CTRL-F model to address the challenges of insufficient spatial modeling, high computational complexity, and limited ability to express position relationships in plant pest image recognition. The rectangular self-calibration module PRCM based on partial channel convolution is designed, and the rectangular self-calibration module RCM and partial channel convolution PConv are combined. This module enhances the recognition ability of different shape targets and reduces the calculation amount while maintaining a strong feature extraction effect. In addition, this paper also replaces the fixed position information representation in the original Transformer with a more flexible relative position coding RPE so that the model can understand the position relationship of objects more accurately when facing images with different sizes or scales. Experimental results on the Plant Village tomato pest image dataset demonstrate that the proposed model achieves an accuracy of 99.83%, surpassing the original model and several mainstream comparison methods.

REFERENCES

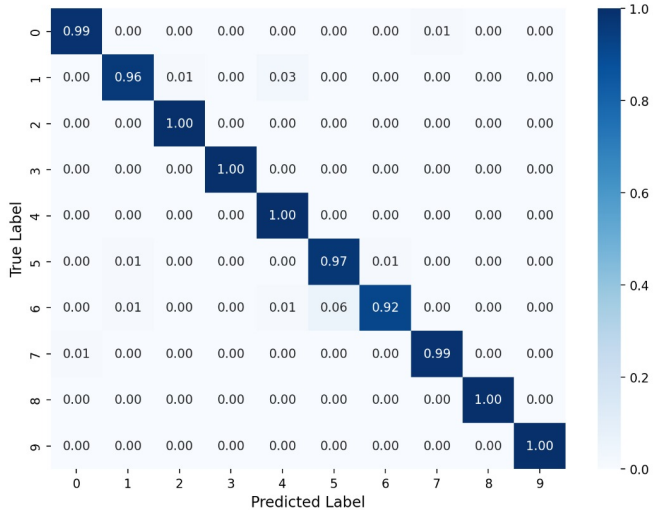
[1] A. Kumar, S. Rani, K. D. Kumar, and M. Jain, *Handbook of AI in engineering applications: tools, techniques, and algorithms*. CRC Press, 2025.
[2] Y. Peng and Y. Wang, "Cnn and transformer framework for insect pest classification," *Ecological Informatics*, vol. 72, p. 101846, 2022.
[3] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019.
[4] S. Praveen and Y. Jung, "Cbam-stn-tps-yolo: enhancing agricultural object detection through spatially adaptive attention mechanisms," 2025. [Online]. Available: https://arxiv.org/abs/2506.07357
[5] H. T. Ung, H. Q. Ung, and B. T. Nguyen, "An efficient insect pest classification using multiple convolutional neural network based models," 2021. [Online]. Available: https://arxiv.org/abs/2107.12189
[6] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7436–7456, 2021.
[7] P. Zhang, G. Zhang, and K. Yang, "Apnet: accurate positioning deformable convolution for uav image object detection," *IEEE Latin America Transactions*, vol. 22, no. 4, pp. 304–311, 2024.
[8] L. Zhang, X. Zheng, J. Ma, S. Mo, and F. Peng, "A hybrid deep learning model for detecting and classifying debris flows and landslides in high-precision aerial images," *Engineering Letters*, vol. 33, no. 9, pp. 3335–3344, 2025.
[9] M. Xu, S. Yoon, Y. Jeong, J. Lee, and D. S. Park, "Transfer learning with self-supervised vision transformer for large-scale plant identification," in *CLEF (Working Notes)*, 2022, pp. 2238–2252.
[10] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: unifying convolution and self-attention for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 581–12 600, 2023.
[11] L. Papa, P. Russo, and I. Amerini, "Meter: a mobile vision transformer architecture for monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5882–5893, 2023.



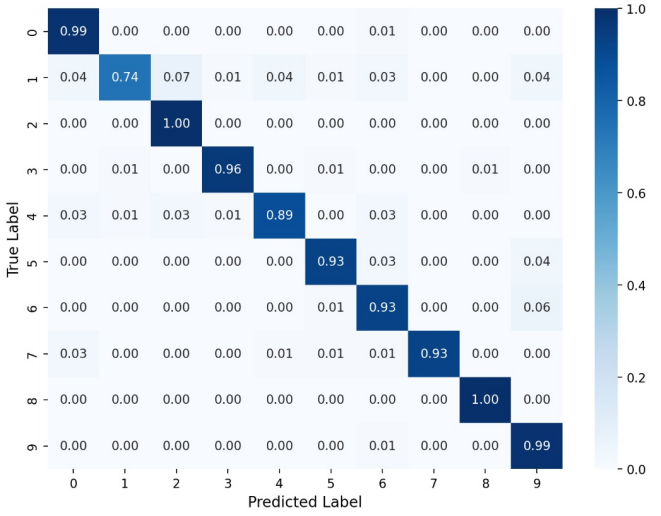
(a) ResNet50



(b) ResNet101



(c) EfficientNetV2-S



(d) ViT-B/16

Fig. 6. Comparison of different models using eight architectures (part 1)

[12] C. Xing, R. Xie, and G. D. Bader, "Retina: reconstruction-based pre-trained enhanced transunet for electron microscopy segmentation on the cem500k dataset," *PLOS Computational Biology*, vol. 21, no. 5, p. e1013115, 2025.

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[14] J. Fan, B. Gao, Q. Ge, Y. Ran, J. Zhang, and H. Chu, "Segtransconv: transformer and cnn hybrid method for real-time semantic segmentation of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1586–1601, 2023.

[15] H. S. EL-Assiouti, H. El-Saadawy, M. N. Al-Berry, and M. F. Tolba, "Ctrl-f: pairing convolution with transformer for image classification via multi-level feature cross-attention and representation learning fusion," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111076, 2025.

[16] J. Guan, R. Lai, Y. Lu, Y. Li, H. Li, L. Feng, Y. Yang, and L. Gu, "Memory-efficient deformable convolution based joint denoising and demosaicing for uhd images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7346–7358, 2022.

[17] Z. Huang, J. Sun, X. Guo, and M. Shang, "One-for-all: an efficient variable convolution neural network for in-loop filter of vvc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2342–2355, 2021.

[18] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 009–12 019.

[19] Q. Guo, X.-J. Wu, T. Xu, T. Si, C. Hu, and J. Tian, "Selective depth attention networks for adaptive multi-scale feature representation," *IEEE Transactions on Artificial Intelligence*, 2024.

[20] Z. Ni, X. Chen, Y. Zhai, Y. Tang, and Y. Wang, "Context-guided spatial feature reconstruction for efficient semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–255.

[21] D. Wang, J. Peng, S. Lan, and W. Fan, "Ctdd-yolo: a lightweight detection algorithm for tiny defects on tile surfaces," *Electronics*, vol. 13, no. 19, p. 3931, 2024.

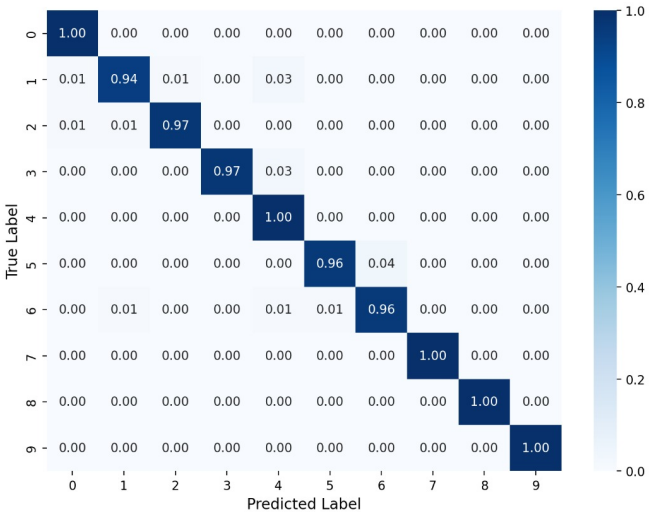
[22] P. Li, M. Wang, Z. Fan, H. Huang, G. Zhu, and J. Zhuang, "Our-net: a multi-frequency network with octave max unpooling and octave convolution residual block for pavement crack segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[23] B. Sun and X. Cheng, "Smoke detection transformer: an improved real-time detection transformer smoke detection model for early fire warning," *Fire*, vol. 7, no. 12, p. 488, 2024.

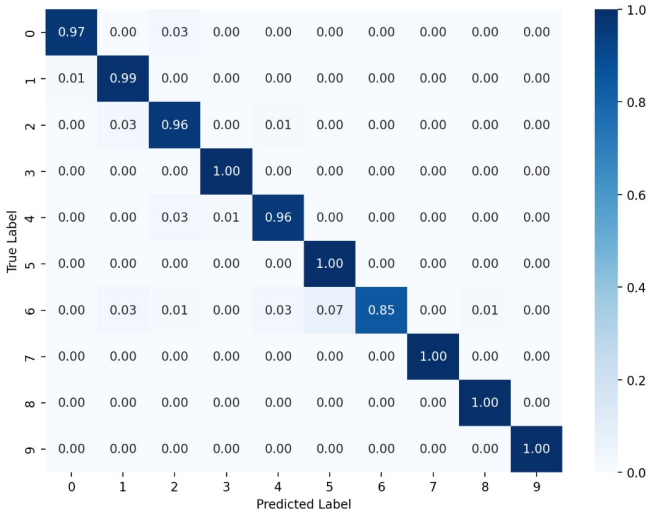
[24] J. K. Kim, D. Park, and M. C. Chang, "Automated risser grade assessment of pelvic bones using deep learning," *Bioengineering*, vol. 12, no. 6, p. 589, 2025.

[25] B. Zhang, Z. Li, B. Li, J. Zhan, S. Deng, and Y. Fang, "Online traffic crash risk inference method using detection transformer and support vector machine optimized by biomimetic algorithm," *Biomimetics*, vol. 9, no. 11, p. 711, 2024.

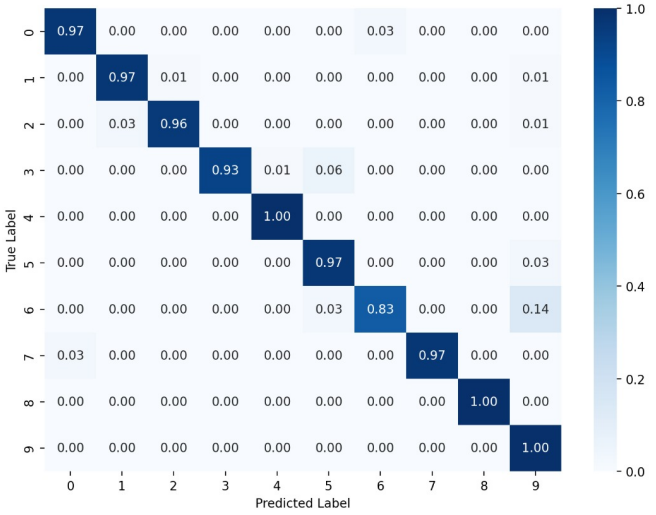
[26] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 021–12 031.



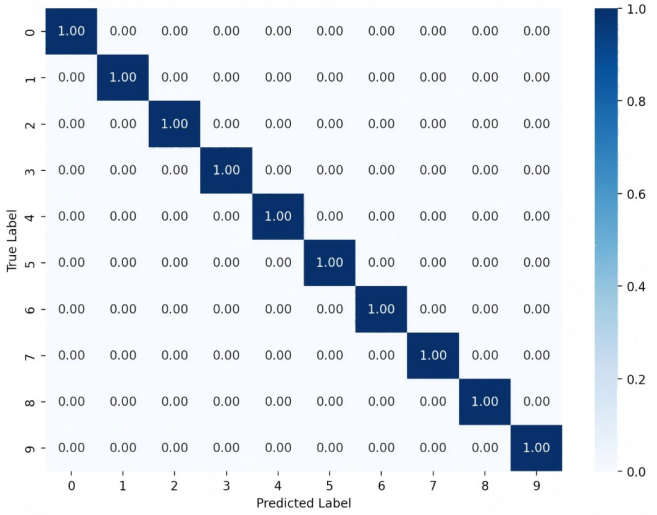
(e) ConvNeXt-T



(f) GoogleNet



(g) Swin Transformer



(h) Proposed model

Fig. 7. Comparison of different models using eight architectures (part 2)

[27] M. Shi, S. Lin, Q. Yi, J. Weng, A. Luo, and Y. Zhou, "Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7401–7416, 2024.

[28] Y. Fang, L. Sun, Y. Zheng, and Z. Wu, "Deformable convolution-enhanced hierarchical transformer with spectral-spatial cluster attention for hyperspectral image classification," *IEEE Transactions on Image Processing*, 2025.

[29] X. Mo, Z. Huang, Y. Xing, and C. Lv, "Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9554–9567, 2022.

[30] J. Zheng, S. Ramasinghe, and S. Lucey, "Rethinking positional encoding," *arXiv preprint arXiv:2107.02561*, 2021.

[31] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, "Toled: tomato leaf disease detection using convolutional neural network," *Procedia Computer Science*, vol. 167, pp. 293–301, 2020.

[32] S. Raschka, "An overview of general performance metrics of binary classifier systems," *arXiv preprint arXiv:1410.5330*, 2014.

[33] L. Gan and Y. Xiao, "Knowledge base question answering based on multi-head attention mechanism and relative position coding," in *Journal of Physics: Conference Series*, vol. 2203, no. 1. IOP Publishing, 2022, p. 012056.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[37] M. Tan and Q. Le, "Efficientnetv2: smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[38] L. Yang, X. Yu, S. Zhang, H. Long, H. Zhang, S. Xu, and Y. Liao, "Googlenet based on residual network and attention mechanism identification of rice leaf diseases," *Computers and Electronics in Agriculture*, vol. 204, p. 107543, 2023.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

Yu-Cheng Li received the B.S. degree in computer science and technology from Taishan University, Taian, Shandong, China, in June 2023. He is currently pursuing a master's degree at the College of Information and Network Engineering, Anhui Science and Technology University, Bengbu, China. His current research direction is mainly applying deep learning methods in plant protection.

Rui Li received the B.S. degree in computer science and technology in 2009 from Anhui University of Science and Technology and the M.S. degree in computer science in 2012 from Anhui University, P.R. China. In 2016, he received his Ph.D. in Applied Computer Technology from Anhui University, P.R. China. He is an associate professor with the College of Information and Network Engineering, Anhui Science and Technology University, P.R. China. His research interests include time-frequency analysis, image processing, and deep neural networks.

Jia-Bing Zhu received his Ph.D. in circuit and system from Anhui University, Hefei, China, in 2008. Currently, he is a professor and PhD supervisor at Huainan Normal University. His research interests include terahertz imaging and satellite communication.