Journal of Agricultural Engineering

https://www.agroengineering.org/

MSIFD-Net: tea stem recognition network based on multi-scale interaction combined feature decoupling

Jiangsheng Gui,[1] Yunxing Huang,[1] Jiancong Wu,[1] Jianneng Chen,[2] Jiangming Jia[2]

[1]School of Computer Science and Technology;
[2]School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China

Corresponding author: Jiangsheng Gui, School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China. E-mail: jsgui@zstu.edu.cn

Publisher's Disclaimer

E-publishing ahead of print is increasingly important for the rapid dissemination of science. The *Early Access* service lets users access peer-reviewed articles well before print/regular issue publication, significantly reducing the time it takes for critical findings to reach the research community.
These articles are searchable and citable by their DOI (Digital Object Identifier).

Our Journal is, therefore, e-publishing PDF files of an early version of manuscripts that undergone a regular peer review and have been accepted for publication, but have not been through the typesetting, pagination and proofreading processes, which may lead to differences between this version and the final one.
The final version of the manuscript will then appear on a regular issue of the journal.

Please cite this article as doi: 10.4081/jae.2026.2317

©The Author(s), 2026
Licensee PAGEPress, Italy

Submitted: 28 May 2026
Accepted: 20 October 2026

***Note**: The publisher is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries should be directed to the corresponding author for the article.*

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

MSIFD-Net: tea stem recognition network based on multi-scale interaction combined feature decoupling

Jiangsheng Gui,[1] Yunxing Huang,[1] Jiancong Wu,[1] Jianneng Chen,[2] Jiangming Jia[2]

[1]School of Computer Science and Technology;
[2]School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China

Corresponding author: Jiangsheng Gui, School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China. E-mail: jsgui@zstu.edu.cn

Data Availability Statement
The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests
The authors declare that they have no competing interests.

Contributions
Jiangsheng Gui: conceptualization, methodology, formal analysis, supervision, investigation, project administration, writing – review & editing, funding acquisition, resources. **Yunxing Huang**: conceptualization, formal analysis, investigation, data curation, writing – original draft, writing – review & editing. **Jiancong Wu**: methodology, investigation, formal analysis, data curation, software, validation, visualization, writing – original draft. **Jianneng Chen**: resources, supervision, data curation, project administration, funding acquisition. **Jiangming Jia**: resources, supervision, data curation, project administration, funding acquisition.

Funding
This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFD2000600), and the National Natural Science Foundation of China (Grant Nos. U23A20175 and 32472009).

Acknowledgments
This work was supported in part by the National Key Research and Development Program of China (Grant No. 2022YFD2000600), the National Natural Science Foundation of China (Grant Nos. U23A20175 and 32472009); Supported by the earmarked fund for CARS-19.
Declaration of generative artificial intelligence and artificial intelligence-assisted

technologies in the writing process
Not applicable.

Abstract

Accurate tea bud detection is essential for intelligent tea-picking robots. After multiple harvests, tea stems appear more frequently, and many tea bud detection models misclassify stems as buds due to subtle differences in shape, texture, and color. Such errors reduce detection accuracy and may damage the end effector when rigid stems are mistakenly selected. In order to reduce the false identification of tea stems as buds, this study proposes a tea stem recognition network based on multi-scale interaction combined with feature decoupling (MSIFD-Net), which operates at the output of a tea bud detection model. The network exploits shallow and deep features to enhance fine-grained discrimination. An Adaptive Channel–Spatial Attention module suppresses irrelevant responses in multi-scale features and emphasizes key spatial regions and channels. A multi-scale interaction structure extracts complementary information from global features across different scales. In addition, a feature decoupling module separates feature components into independent subspaces to reduce redundancy. On the ZC108 dataset, the integration of MSIFD-Net into YOLOv11 achieved a peak AP of 91.28%, improving by 10.23 percentage points compared with YOLOv11 without MSIFD-Net. Furthermore, the FDR decreased markedly from 22.22% to 1.1%, representing a reduction of 21.12 percentage points. These results demonstrated that the proposed method effectively reduced false detections caused by residual tea stems and decreased the risk of unnecessary picking operations and potential damage to the tea-picking robot end effector.

Key words: attention mechanism; feature decoupling; multiple harvest; multi-scale interaction; tea stem recognition.

Introduction

Tea is one of the three major beverages consumed worldwide and plays an important role in agricultural economies and cultural traditions across many countries (Kumar *et al*., 2011). In particular, high-grade teas characterized by the “one leaf and one bud” (1L1B) or single-bud

standard are highly valued due to their superior quality and strong market demand. However, manual tea harvesting suffers from low efficiency and high cost, while traditional mechanical tea harvesting lacks selectivity and shows poor performance in bud collection. This severely restricts the high-quality development of the premium tea industry. Therefore, there is an urgent need for research and development of automated, intelligent tea-harvesting technology.

Recent advances in artificial intelligence and computer vision have greatly promoted deep learning-based tea bud detection. Early studies applied conventional detection models: Chen and Chen (2020) combined RCNN with FCN for bud localization, Wang *et al*. (2023b) used Mask-RCNN with feature aggregation for detection and segmentation, and Xie and Sun (2023) proposed Tea-YOLOv8s to achieve better performance. Later works added deformable convolution, global attention, and enhanced spatial pyramid pooling, together with data augmentation, which increased accuracy.

Small, occluded, or unstructured tea buds present a major challenge and cause high rates of false positives and false negatives. Wang *et al*. (2023a) proposed TBC-YOLOv7 with Transformer-based feature enhancement, which reached 87.5% mAP but raised computational cost. Lightweight models provide a balance between efficiency and accuracy: Guo *et al*. (2023) combined YOLOv4-lighted with ICBAM, reduced model size by 75%, and increased speed; Gui *et al*. (2023) replaced standard convolutions with Gaussian convolutions and applied CIoU for accurate detection; Yang *et al*. (2024) developed a YOLOX-based model able to recognize multiple tea bud varieties.

The deep learning-based tea bud detection methods described above have achieved superior performance. Nevertheless, these studies mainly focused on tea bud detection during the first harvest and rarely considered the repeated harvesting operations required in practical tea production. After repeated harvesting, residual tea stems remain at previously picked positions, and their green appearance and shape become highly similar to those of harvestable buds. Under such conditions, existing bud detection models may incorrectly identify residual stems as buds,

as shown in 

Figure 1. When these false targets are selected for robotic picking, the resulting operations fail to produce harvestable buds, thereby consuming additional picking cycles and reducing the efficiency and continuity of harvesting operations. In some cases, repeated invalid interactions may also increase the mechanical burden on the end effector.

Therefore, a tea stem recognition network based on multi-scale interaction combined with feature decoupling (MSIFD-Net) was proposed and integrated into the output stage of tea bud detection networks to further improve the reliability of picking target selection after multiple harvesting operations. This network decoupled the visual feature representations of tea stems and buds, enabling more accurate discrimination of subtle differences between them and reducing false detections caused by residual stems. The contributions of this study were summarized as follows:

- An Adaptive Channel–Spatial Attention (ACSA) module was designed to enable adaptive weight adjustment of multi-scale feature maps, which enriched contextual information and enhanced visual representation.

- A multi-scale interaction (MSI) structure was proposed to update the shallow features with the prior knowledge obtained from the deep features. The structure obtained an attention region map rich in representational information. Then, it fused all attention regions to reduce noise information and enhanced the original input image.

- A feature decoupling (FD) module was integrated into the feature extraction network to

disentangle feature components according to their relevance to bud–stem classification and reduce redundant information through a soft orthogonality constraint.

Materials and Methods

Data collection

The collection of tea images represents a crucial first step in the task of tea stem recognition. In this study, a tea dataset named Zhongcha 108 (ZC108) was collected in Hangzhou, Zhejiang, China, in March 2024. The dataset contained 500 images. All images were acquired from tea plants of the Zhongcha 108 cultivar. The dataset covered tender young shoots ready for harvesting, including unopened buds and early-unfolding young leaves, consistent with the picking standard used at the collection site. The images were captured after repeated tea bud harvesting to include residual tea stems alongside young shoots. The image samples of ZC108 dataset were shown in **Figure 2**.

The images were captured outdoors under natural illumination on sunny days. Appropriate shading was applied during image acquisition to avoid intense direct sunlight and backlighting. The capture device was a Zivid 2 3D camera. The size of the images was fixed at 4032×2268 pixels, and the storage format was JPG. During the image capture phase, the capture angle was about 45 degrees, and the capture height was about 35 cm. The images were captured after multiple tea bud harvests to ensure a sufficient number of tea stems for model training. Only RGB images were used as model inputs.

Subsequently, target regions within the images were categorized into tea buds and tea stems. These regions were then cropped and saved as independent instances. The numbers of instances of the two were 9,670 and 2,984, respectively. Some instances of tea buds and tea stems were shown in **Figure 3**.

Finally, the dataset was partitioned based on the original images. Specifically, 100 images (with 2,638 instances) were reserved as an independent test set. Out of the other 400 images, a division into train and val sets was made at an 8:2 ratio, which consisted of 7,680 and 2,336 instances, respectively. The specific data split is detailed in

Table 1.

Model introduction

MSIFD-Net was a CNN-based binary classifier that verified whether buds detected by detection models were actual tea buds or tea stems. Detections identified as stems were discarded, while true buds were retained. The overall architecture of MSIFD-Net was shown in **Figure 4**. In this architecture, the bottom-left part denoted the input of the recognition model, and the input images were cropped from the bounding boxes generated by the tea bud detection model.

First, the input images passed through a five-stage CNN backbone, with stages referred to as Stage 1 to Stage 5. Multi-scale feature maps were produced after Stage 3, Stage 4, and Stage 5, with spatial resolutions of 56×56, 28×28, and 14×14 from shallow to deep layers, respectively. Second, the Adaptive Channel–Spatial Attention (ACSA) module refined intermediate multi-scale features by adaptively emphasizing channels and spatial regions, which enhanced discriminative local responses. These refined features were then delivered to the middle multi-scale interaction (MSI) structure.

Within the MSI structure, the multi-scale features refined by ACSA underwent further enhancement through basic convolution (BC) modules, which emphasized regions that might contain critical discriminative cues. Subsequently, feature maps that contained salient regional information at different scales were cropped and upsampled to the input resolution of 448×448, which yielded three salient feature maps, referred to as attention images. Afterward, a single concentrated feature map, termed the highlight image, was obtained through feature fusion across the three scales. This highlight image was fed back to the backbone input and was used in subsequent optimization stages, which further enhanced feature representation for tea bud and tea stem discrimination.

Finally, the feature decoupling (FD) module removed redundant information related to tea stem recognition and retained only task-relevant features that contributed effectively to the final classification decision.

Adaptive Channel-Spatial Attention module

In order to enhance feature representation under complex spatial variations, an Adaptive Channel–Spatial Attention (ACSA) module was introduced to refine intermediate feature maps. Unlike conventional attention mechanisms that modeled channel attention and spatial attention independently or in a fixed sequential order, ACSA established bidirectional interaction

between channel semantics and spatial distributions. This design enabled joint optimization of global semantic importance and fine spatial localization, which proved effective for cluttered natural scenes.Given an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$, the ACSA module refined features through two interdependent processes: spatially-conditioned channel attention and channel-guided spatial attention. The overall transformation of ACSA was expressed as

$$\hat{F} = M_s \odot (M_c \odot F) \tag{1}$$

where M_c and M_s denoted the channel attention map and spatial attention map, respectively. The schematic architecture of the Adaptive Channel–Spatial Attention module was illustrated in

Figure 5.

Spatially-conditioned Channel Attention: Conventional channel attention mechanisms typically relied on global statistical descriptors and assumed that channel importance is independent of spatial distribution. However, under complex spatial patterns, discriminative channel responses often corresponded to localized structural cues, which made channel importance strongly dependent on spatial activation. To address this limitation, a spatial conditioning strategy was adopted, which allowed channel attention to adapt to spatial context.

Specifically, spatial dependency was encoded through a lightweight depthwise convolution that preserved channel specificity while it captured local spatial context:

$$\Phi_S = \sigma\big(DwConv_{7\times 7}(F)\big) \tag{2}$$

where Φ_S represented the spatial conditioning feature. The spatial response was combined with the input feature to form a spatial aware channel descriptor. Adaptive channel attention was computed using this descriptor together with global statistical descriptors as follows:

$$M_c = \sigma\left(MLP\big(GAP(F)\big) + MLP\big(GMP(F)\big) + MLP\big(GAP(\Phi_S \odot F)\big)\right) \tag{3}$$

The input feature was reweighted along the channel dimension:

$$F_c = M_c \odot F \tag{4}$$

Through spatial conditioning, channel importance was no longer determined solely by global statistics but was modulated according to spatial activation.

Channel-guided Spatial Attention: Although spatial attention aimed to emphasize informative regions, direct spatial estimation without channel semantic awareness often led to inaccurate

emphasis. To alleviate this issue, channel-guided spatial attention was employed, in which spatial response was conditioned on channel-refined features.

Since F_c already reflected channel semantic emphasis, spatial attention was derived from channel-refined responses. Channel-wise statistical descriptors were first obtained:

$$F_s^{pool} = Concat\big(AvgPool_c(F_c), MaxPool_c(F_c)\big) \tag{5}$$

where $AvgPool_c(\cdot)$ and $MaxPool_c(\cdot)$ denoted average and max operators along the channel axis. The spatial attention map was then produced as

$$M_s = \sigma\left(Conv_{7\times7}\left(F_s^{pool}\right)\right) \tag{6}$$

Finally, the spatial attention map was applied to the channel-refined feature:

$$\hat{F} = M_s \odot F_c \tag{7}$$

By derivation of spatial attention from channel-refined features, spatial emphasis was guided by channel semantics, which resulted in more accurate localization of informative regions.

The ACSA module achieved effective feature refinement through bidirectional cross-dimensional interaction, where spatial context affected channel importance and channel semantics guided spatial emphasis. The module introduced limited computational cost, as it mainly relied on depthwise convolution and lightweight statistical operators, and maintained full compatibility with standard convolutional backbones.

Multi-scale interaction structure

The multi-scale interaction (MSI) structure was designed to progressively extract and exploit image information at multiple scales through a staged strategy, which improved classification accuracy. The overall architecture of MSI was illustrated in **Figure 6.** Feature maps from three stages (Stage 3 to Stage 5) were first enhanced by the ACSA module to emphasize discriminative information. Within the MSI framework, these enhanced feature maps were then passed to the basic convolution (BC) feature refinement module, which re-extracted and further processed feature information at different scales. After this operation, class activation maps (CAM) were employed to perform a second-stage crop, and regions of interest at multiple scales were fused (Zhou *et al*., 2016).

The BC module mainly performed further feature refinement as follows: first, it adjusted the

number of channels to a uniform value using 1×1 convolution; then it strengthened the spatial features with 3×3 convolution; finally, it introduced non-linearity through the ReLU activation function to refine multi-scale feature information.

After feature refinement, bilinear interpolation upsampled each feature map and produced attention maps that highlighted salient regions in the original image. Based on these attention maps, the height and width of each salient region were computed, and the region boundaries were extended by a predefined expansion ratio. The salient regions were then cropped and resized to the original image resolution, which yielded the final regions of interest for subsequent network inference.

The feature map at each scale was represented as $x_n \in \mathbb{R}^{C_n \times H_n \times W_n}$, where H_n, W_n, and C_n represented the height, width, and number of channels respectively. A set of functions $\{F_1(.), F_2(.), \ldots, F_n(.)\}$ was used to compress $\{x_1, x_2, \ldots, x_n\}$ into one-dimensional vector descriptors $\{v_1, v_2, \ldots, v_n\}$, and $v_n \in \mathbb{R}^{Cv}$. Cv represented the length of the one-dimensional vector descriptor. The length of the one-dimensional vector descriptor given by the multi-scale network was the same. The function $F_n(.)$ for processing x_n was defined as:

$$V_n = K f^{GMP}\left(f^{BC}(x_n)\right) \tag{8}$$

Where f^{GMP} denoted global max pooling. f^{BC} denoted basic convolution. x_n was the feature map generated by the network. It was assumed that there was a total of k classes, and $k_n \in \{1,2,\ldots.,K\}$ was the category predicted by the network. From the inference result of the network, the class activation map (CAM) of x_n could be obtained. The CAM of a specific class specified the discriminative image regions used by the CNN to identify this class. It helped to determine the regions in the image that played a key role in the recognition of a specific class through visualization. It provided important interpretive information that helped the model analyze and understand the relationship between image features and categories, and assisted in the assessment and improvement of the network's ability to recognize targets of different categories. The CAM generated by the network was defined as:

$$L^n(x,y) = \sum_k \omega_k^n f_k(x,y) \tag{9}$$

Where $L^n(x,y)$ was the value of the activation map of category n at position (x,y), ω_k^n was the weight of the k-th feature map of category n, and $f_k(x,y)$ was the value of the k-th feature map

at position (*x,y*). $L^n(x, y)$ directly represented the importance of activation at the spatial position (*x,y*), which guided the network to classify the image as class k_n.

Feature decoupling module

Inspired by shared-private learning (Bousmalis *et al*., 2016) and pose and expression disentanglement in video portrait editing (Salzmann *et al*., 2010; Song *et al*., 2012), this study integrated a feature decoupling module related to classification attributes into the feature extraction backbone. The module separated different features in the image into different feature subspaces to eliminate redundant features. The essence of this disentangled representation lay in the construction of orthogonal feature subspaces that separately modeled semantic attributes directly related to class discrimination and scene-independent interference factors. This method adopted a two-branch network structure to achieve feature space separation. The class-relevant branch focused on discriminative morphological features of tea buds and tea stems, whereas the class-irrelevant branch isolated interference factors such as lighting variations and occlusion. Finally, through feature projection, the two types of attributes were mapped into orthogonal latent spaces to eliminate the redundant coupling among the original features.

The specific structure of the FD module was illustrated in **Figure 7**. The input feature map F was projected and decomposed into two orthogonal latent subspaces. A soft orthogonality constraint between the two representations was imposed to ensure non-redundancy (Liu *et al*., 2017; Ruder and Plank, 2018). Specifically, the input feature map F was first convolved and flattened, after which multiple multi-layer perceptron (MLP) layers were employed to decouple the latent space into two orthogonal subspaces, namely the category-related space and the category-unrelated space. Two different sets of convolutional layers were used as the decoupling backbone, and the decoupled features were obtained after they pass through several shared MLP layers. The decoupling process was expressed as:

$$c, \bar{c} = MLP\big(Conv(F)\big)cls \tag{10}$$

where c and $\bar{c}$ represented two new feature vectors obtained after the feature decoupling operation, and both were used in the subsequent classification head computation to obtain the classification loss.

Difference loss (L_{diff}). This loss was designed to ensure that the category-related and category-unrelated representations captured different aspects of the input features respectively. Within a training batch, let $\mathrm{H_F^c}$ and $\mathrm{H_F^{\bar{c}}}$ denote two matrices whose rows corresponded to the hidden vectors $\mathrm{h_F^c}$ and $\mathrm{h_F^{\bar{c}}}$ derived from the input feature map F. The orthogonality constraint between the two feature representations was calculated as:

$$\left\|\mathrm{H}_F^{c\,T}\mathrm{H}_F^{\bar{c}}\right\|_F^2 \tag{11}$$

where $\|*\|_{\mathrm{F}}^2$ denoted the squared Frobenius norm. The overall difference loss was defined as:

$$L_{diff} = \sum \left\|H_F^{c\,T} H_F^{\bar{c}}\right\|_F^2 \tag{12}$$

Classification loss (L_{cls}). This loss was used to measure the accuracy of the model's predictions for each object category. A category label was assigned to each object to be detected, and the model was required to predict the category of the object within each bounding box. The classification loss L_{cls} quantified the discrepancy between the predicted category and the ground-truth label. Specifically, given the ground-truth label Y, the predicted output $\widehat{\mathrm{Y}}$, and the total number of categories n, L_{cls} was formulated as:

$$L_{cls} = -\sum_{i=1}^{n} y_i log\widehat{y_i} \tag{13}$$

Experiments results and analysis

Experimental configurations

The experiments in this study were conducted on a dedicated deep learning server. The software configurations consisted of the Ubuntu 20.04 operating system, the Python 3.9.18 programming language, the PyTorch 2.1.2 development framework, the CUDA 12.1 computing platform, and the Visual Studio Code 2025 IDE. The hardware configurations consisted of an Intel Xeon Silver 4310 CPU with a clock speed of 2.10 GHz, 256 GB of RAM, and NVIDIA GeForce RTX 3090 GPU with 24 GB of video memory.

During the training process, stochastic gradient descent (SGD) was adopted as an optimization strategy, with a momentum of 0.9 and a weight decay factor of 0.0005. The initial learning rate was 0.001, and a cosine annealing schedule was applied.

Evaluation metrics

The effectiveness of the proposed method was assessed through a set of established metrics. These metrics included parameters (Params), floating-point operations (FLOPs), accuracy (ACC), precision (P), recall (R), F1 score, false detection rate (FDR) and Average Precision (AP). Among these, FDR and AP were used to evaluate the tea bud detection model before and after the stem recognition model was introduced, and the other metrics were used for the stem recognition model itself. Specifically, Params reflected the scale of the model, and FLOPS was used to evaluate the computational cost and detection speed of the model. Accuracy represented the ratio of all correctly identified samples to the total number of samples. Precision referred to the ratio of the correctly identified positive samples to the total number of predicted positive samples. Recall referred to the ratio of the correctly identified positive samples to the total number of actual positive samples.

In tea bud detection tasks, FDR directly reflected the ratio of stem samples misidentified as bud samples to the total number of predicted bud samples, and the average precision was introduced to establish an evaluation metric based on confidence value ranges for image retrieval. In addition, true positive (TP) represented the number of samples predicted as positive and actually positive. False positive (FP) represented the number of samples predicted as positive but actually negative. True negative (TN) and False negative (FN) were defined in a similar manner. The formulas were defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$FDR = \frac{FP}{TP + FP} \tag{17}$$

$$AP = \int_0^1 P(R)\, dR \tag{18}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

Results and analysis

Comparison of different backbone networks

To explore the accuracy of tea stem recognition using different feature extraction networks,

three representative backbone networks in terms of performance and lightweight design, namely ShuffleNet V2 (Ma *et al*., 2018), ResNet18 (He *et al*., 2016), and MobileNet V3 (Howard *et al*., 2019), were selected as backbones for MSIFD-Net in comparative experiments. ShuffleNet V2 offered four channel size configurations, which enabled researchers to explore the trade-off between model performance and parameter scale from multiple dimensions such as model effectiveness and the number of parameters. As a classic deep residual network, ResNet18 was widely adopted in image recognition tasks and demonstrated strong representational capability, which offered a rigorous benchmark for the evaluation of lightweight network architectures. MobileNet V3 was available in two variants, large and small, and these variants effectively demonstrated how lightweight architectural design contributed to improvements in inference speed. The specific experimental results were shown in

Table 2.

As can be seen from the comparison results in

Table 2, ShuffleNet V2 demonstrated obvious advantages over ResNet18 and MobileNet V3. In terms of parameters, the ShuffleNet V2 2.0× model had a parameter count of 5.42M, which was lower than ResNet18's 11.24M. However, in terms of accuracy (ACC), it reached 93.85%, which was higher than the values for MobileNet V3 Large (91.24%) and ResNet18 (92.10%). The model achieved high accuracy with relatively few parameters, and this characteristic demonstrated the advantage of its architecture in balancing model complexity and performance.

Comparison of different attention mechanisms

In MSIFD-Net, the ACSA module adaptively allocated weights to multi-scale features across multiple dimensions. In order to verify the effectiveness of this attention mechanism, comparative experiments were conducted compare the model without attention, the model with ACSA, and four other mainstream attention mechanisms: SE (Hu *et al*., 2018), CBAM (Woo *et al*., 2018), GAM (Liu *et al*., 2021), and CPCA (Huang *et al*., 2024). The lightweight backbone ShuffleNetV2 2.0× was used as backbone in this experiment. The experimental results were detailed in

Table 3. These models were evaluated using the number of parameters and accuracy.

The experimental results showed that the SE mechanism, which only establishes interdependencies between feature channels, provided limited enhancement to model performance. In contrast, the CBAM mechanism incorporated additional spatial attention and yielded better recognition results for tea stems, yet precision remained limited to 96.60%. A comparison among GAM, CPCA, and the baseline indicated that these methods did not reach optimal performance levels in tea bud detection tasks.

In comparison, the proposed ACSA attention mechanism achieved a significant performance

enhancement, with a 3.50% increase in precision and a 2.83% increase in ACC over the baseline. Specifically, ACSA reached a precision of 98.40% and an ACC of 96.68%, both of which outperformed CBAM, GAM, and CPCA. This superiority arises from its unique design and precise adaptation to task requirements. ACSA captures subtle feature differences between tea stems and buds in finer detail through an innovative global attention interaction method. Meanwhile, its structural design achieves an efficient balance between computational cost and performance (5.83 M). Consequently, the model reaches higher accuracy at a low computational cost. This precise feature capture capability and balance between cost and effectiveness makes ACSA significantly superior to CBAM, GAM, and CPCA in tea stem recognition tasks.

Comparison of multi-scale interaction structure at different scales

This section primarily investigated the performance of the MSI structure at different scales. The ShuffleNetV2 2.0× model with the best performance was selected as the baseline to compare the performance of two-scale, three-scale, four-scale, and five-scale network structures. The purpose of the experiment was to verify which scale design could most effectively balance the model's ability to learn and its generalization ability. The impact of scale on model performance was assessed by varying only the number of scales, and all other network parameters and training conditions stayed the same.

From the experimental results in **Table 4**, it could be concluded that the three-scale MSI with three-scale demonstrated the best performance. These results not only confirmed that multi-scale networks with an appropriate number of scales could more effectively extract and utilize multi-scale features in the data, but also revealed a non-linear relationship between the number of scales in multi-scale networks and network performance: too many scales (i.e., more than three) may have caused the model to focus excessively on shallow features and neglect deeper abstract information, which thereby limited the model's generalization ability. Conversely, multi-scale networks with too few scales may have failed to fully capture the complexity and diversity of the data, which resulted in poor performance on unseen data and impaired generalization ability as well.

Visualization analysis of multi-scale interaction structure

To further explore the ability of the three-scale MSI to focus on the local receptive fields of targets, a detailed visualization analysis was conducted, as was shown in **Figure 8**, where (a), (b), (c), and (d) represented two-scale, three-scale, four-scale, and five-scale networks, which corresponded to the serial numbers in **Table 4**. The test samples included three groups each of tea buds and tea stems. Through gradient-weighted class activation mapping (Grad-CAM) heatmaps, the image regions focused on by the model could be visually examined. In the heatmaps generated by Grad-CAM, different colors represented different attention levels of the model: red regions indicated the highest attention; orange and yellow regions indicated slightly high attention; green regions indicated slightly low attention; blue regions indicated low attention.

An examination of these heatmaps revealed significant differences in the feature extraction of tea buds and tea stems across different scales were evident. Specifically, in the two-scale network, although the heatmap captured some key features, the overall receptive field was relatively scattered and less concentrated. In the three-scale network, the heatmap exhibited a more complete and target-aligned visual receptive field. Notably, when the model recognized buds, it primarily focused on the central regions and edges of the leaves, which typically had a fresh green color and prominent contours; when the model recognized tea stems, it paid more attention to the rough texture of the stems and obvious cut cross-sections. This attention distribution not only aligned with the actual features of tea buds and tea stems but also further validated the accuracy and efficiency of the three-scale network in feature extraction.

In contrast, although the four-scale and five-scale networks performed feature extraction at deeper scales, their heatmaps did not present better results. Instead, these networks often over-focused on details or noise during feature extraction, and this resulted in incomplete and inaccurate overall receptive fields. This explained why the three-scale structure provided the greatest improvement in model performance in the experimental results. Through a comparison of heatmaps from networks with different numbers of scales, the advantages of the three-scale network in the extraction of features of tea buds and tea stems became more visually apparent, and these advantages appeared were reflected not only in the integrity and accuracy of the receptive fields but also in the model's sensitivity and robustness to target features. Therefore, it can be concluded that the three-scale network structure significantly enhances the

performance of MSIFD-Net.

Ablation study

The robustness of each proposed module, including FD, ACSA, and MSI, as well as the effect of each optimization strategy on model performance, was evaluated through ablation experiments based on the ShuffleNet V2 backbone. Different module combinations were added to verify the contribution of each component. Here, "√" indicates that the optimization module in question has been added, and "-" indicates that it has not been added. In this experiment, the mixed tea dataset was used uniformly for training, and all other training parameter configurations were consistent. The results of the ablation experiment were shown in

Table 5.

As shown in

Table 5, the introduction of each proposed module led to a consistent improvement in model performance on the ShuffleNet V2 backbone. When only the FD module was incorporated, the accuracy increased from 93.85% to 94.62%, accompanied by gains in precision and F1 score. Further performance improvement was observed after the addition of the ACSA module. The accuracy reached 97.12% and the precision rose to 98.26%. After all modules were integrated, the complete MSIFD-Net achieved the best performance, with an accuracy of 97.35%, a precision of 98.92%, and an F1 score of 0.975.

These results demonstrated that each module provided effective performance enhancement and that the combination of FD, ACSA, and MSI produced the best overall result. The ablation results confirmed that the proposed modules played complementary roles and improved model capability for tea bud detection tasks. In practical applications, an appropriate combination of backbone networks and innovative modules can be chosen according to specific requirements and resource constraints to achieve optimal performance. Moreover, these experimental results provide a valuable reference for further research and improvement of the model.

Comparison of different detection models before and after adding MSIFD-Net

To evaluate bud–stem discrimination in detected tea bud candidates under repeated harvesting conditions, MSIFD-Net was integrated into the output stage of each detection model. Image regions cropped from the predicted bounding boxes were fed into MSIFD-Net to classify each candidate as a tea bud or tea stem. Candidates identified as tea stems were removed, while tea

bud candidates were retained to refine the final detection results.

YOLOX was selected as an anchor-free detector with a decoupled head (Ge *et al*., 2021), while YOLOv8–v11 were included to evaluate the effectiveness of MSIFD-Net across different YOLO architectures. RT-DETR was further introduced as a Transformer-based real-time end-to-end detector (Zhao *et al*., 2024). These detectors represented different architectural paradigms and provided diverse baselines for evaluating the effectiveness of candidate verification. Paired comparisons before and after integrating MSIFD-Net were conducted to determine whether stem-related false detections were reduced across different detection frameworks.

AP, FDR, and FPS were compared before and after integrating MSIFD-Net. The evaluated models included YOLOX, YOLOv8, RT-DETR, YOLOv9, YOLOv10, and YOLOv11. The 500 original images from the ZC108 dataset were used to construct the training, validation, and test sets. The experimental results are presented in

Table 6.

The results showed that MSIFD-Net consistently reduced FDR and improved AP across all evaluated detectors. Specifically, for YOLOv11, integrating MSIFD-Net reduced FDR from 22.22% to 1.10% and increased AP from 81.05% to 91.28%, corresponding to improvements of 21.12 percentage points and 10.23 percentage points, respectively. YOLOv11+MSIFD-Net achieved the highest AP among all evaluated detector–classifier combinations. These results demonstrated that MSIFD-Net effectively suppressed stem-induced false detections and enhanced the robustness of tea bud detection under repeated harvesting conditions.

Visualization result

In this section, an example of the filtering effect of MSIFD-Net on the tea bud candidates detected by the detection model was shown in **Figure 9**. The results in group (a) were the predictions of the detection model, and the results in (b) were obtained after MSIFD-Net filtered them. The red boxes represented the detected tea bud targets, and the blue boxes represented tea stem targets recognized by MSIFD-Net.

The recognition results demonstrated that integrating MSIFD-Net into the output stage of the detection model effectively improved detection precision. Compared with the original detection

results containing false detections of tea stems, MSIFD-Net further optimized the outputs of the detection model. It accurately identified and distinguished tea stem samples, effectively removed false detections, and significantly reduced the FDR. Removing false detections allowed the tea-picking equipment to operate more accurately, resulting in improved stability and a lower error rate during robotic tea picking in field environments.

Discussion

Accurate discrimination between tea buds and tea stems remains a critical challenge in intelligent tea harvesting systems. Visual similarity, occlusion, and structural overlap often cause misclassification in conventional detection frameworks. The results in

Table 6 showed that integrating MSIFD-Net improved AP and reduced FDR across all evaluated detectors. For YOLOv11, AP increased from 81.05% to 91.28%, while FDR decreased from 22.22% to 1.10%, corresponding to improvements of 10.23 and 21.12 percentage points, respectively. These results demonstrated that introducing a separate bud–stem verification stage after detection effectively improved the reliability of candidate regions in the evaluated repeated-harvest scenario.

Related studies have addressed visual ambiguity in tea-harvesting perception from different perspectives. Yang *et al*. (2024) employed YOLOX to distinguish visually similar tea buds among different cultivars, mainly focusing on cultivar-level recognition differences. Guo *et al*. (2023) developed a lightweight YOLOv4-based detector with ICBAM for tea bud detection and further performed segmentation and depth-based filtering within detected regions to achieve picking-point localization. Xie and Sun (2023) enhanced tea bud detection by incorporating deformable convolution, attention mechanisms, and spatial pyramid pooling into the detector architecture. These studies addressed cultivar discrimination, detector enhancement, or depth-assisted picking-point localization. In contrast, MSIFD-Net focused on verifying bud–stem identity within detected candidate regions, adding an additional target-category check before subsequent picking decisions. This verification strategy complemented existing detection and localization approaches by specifically reducing false positives caused by residual tea stems in repeated-harvest scenarios.

The introduction of the ACSA module contributed to adaptive feature recalibration across

channel and spatial dimensions. The attention comparison in

Table 3 supported the effectiveness of this joint refinement strategy for bud–stem classification. The MSI structure facilitated cross-scale information exchange and promoted multi-granularity semantic representation. Among the evaluated scale configurations, the three-scale structure achieved the highest ACC (**Table 4**), indicating that appropriate scale fusion was beneficial for capturing discriminative characteristics of tea buds and stems. Furthermore, Grad-CAM visualization in **Figure 8** showed that model responses were concentrated on leaf regions and stem cut surfaces, providing qualitative evidence that the network captured morphological cues associated with bud–stem discrimination. These results suggested that combining multi-scale feature interaction with adaptive feature refinement improved the representation of subtle structural differences between tea buds and stems.

The FD module encouraged the separation of feature components according to their classification relevance through a soft orthogonality constraint. This design reduced feature coupling between class-related cues and irrelevant variations, thereby improving discrimination reliability. In the staged ablation experiment (

Table 5), ACC increased from 93.85% for the backbone network to 97.35% for the complete model, representing an improvement of 3.50 percentage points. These results supported the contribution of the proposed modules to fine-grained classification performance and further explained the improvements observed in detection-level AP and FDR after integrating MSIFD-Net with YOLOv11.

The improved discrimination performance was accompanied by additional computational costs. As shown in

Table 6, the detectors integrated with MSIFD-Net achieved processing speeds of 92–154 FPS on the experimental platform equipped with an NVIDIA GeForce RTX 3090. For YOLOv11, the processing speed decreased from 195 to 154 FPS after integration, corresponding to a reduction of approximately 21.0%. Despite the computational overhead introduced by candidate verification, all evaluated detectors maintained processing speeds above 90 FPS, demonstrating high processing throughput under the evaluated conditions. These results indicated that the proposed method could satisfy the requirements for real-time visual processing on the experimental platform. However, practical onboard performance depends not only on model inference speed but also on camera input frequency, onboard computing

resources, the number of candidate regions requiring verification, image preprocessing, communication latency, and subsequent picking control. Therefore, further optimization and evaluation on embedded platforms are required to determine whether the complete perception pipeline can meet the update-rate and response-time requirements of robotic tea picking under field conditions.

The proposed verification module evaluated candidate regions generated by the preceding detector. Therefore, its contribution mainly involved rejecting stem candidates while preserving correctly detected buds, whereas recovering missed buds remained dependent on the detection stage. Moreover, practical robotic harvesting requires additional modules, including spatial localization, grasping-point estimation, and motion planning. The improved candidate classification provided by MSIFD-Net could therefore contribute to more reliable target selection within the overall tea-picking workflow.

The dataset consisted of RGB images of a single cultivar, Zhongcha 108, containing harvestable young shoots captured under outdoor illumination conditions with shading strategies to reduce severe illumination variations. However, complex field environments may introduce additional challenges, including extreme lighting changes, dense occlusion, and variations among growth stages. Future work will expand dataset diversity and optimize inference efficiency for embedded tea-picking platforms, particularly considering scenarios involving multiple candidate regions and resource-constrained deployment.

Conclusions

This study proposed a tea stem recognition network named MSIFD-Net, which integrates multi-scale interaction and feature decoupling to improve discrimination between tea buds and tea stems. The proposed network strengthened global semantic representation at multiple granularities and improved feature quality, which enabled accurate tea stem recognition and enhanced the efficiency and durability of intelligent tea-picking equipment.

An Adaptive Channel–Spatial Attention (ACSA) module was introduced to adaptively adjust feature weights for tea buds and tea stems and to enrich contextual information. A multi-scale interaction (MSI) structure was designed to activate latent global features and to construct multi-scale attention regions. In addition, a feature decoupling (FD) module disentangled

classification-relevant and irrelevant feature components and reduced redundant information through a soft orthogonality constraint, thereby improving the discrimination between tea buds and stems.

Experimental results on the ZC108 dataset demonstrated the effectiveness of MSIFD-Net. When integrated into the YOLOv11 detection framework, MSIFD-Net achieved an AP of 91.28%, representing an improvement of 10.23 percentage points over the standalone YOLOv11, while reducing the FDR from 22.22% to 1.10%. The reduction in false stem-to-bud detections improved the reliability of candidate targets for subsequent picking decisions and reduced unnecessary picking operations.

The experiments used RGB images of the Zhongcha 108 cultivar, with harvestable young shoots photographed under sunny outdoor conditions with shading to avoid intense direct sunlight and backlighting. Future work will expand the dataset under complex environmental conditions and will further optimize the network structure to improve real-time performance on embedded intelligent tea-picking systems.

References

Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D, 2016. Domain separation networks. arXiv:1608.06019v1

Chen YT, Chen SF, 2020. Localizing plucking points of tea leaves using deep convolutional neural networks. Comput Electron Agr 171:105298.

Ge Z, Liu S, Wang F, Li Z, Sun J, 2021. YOLOX: Exceeding YOLO series in 2021. arXiv:2107.08430

Gui Z, Chen J, Li Y, Chen Z, Wu C, Dong C, 2023. A lightweight tea bud detection model based on YOLOv5. Comput Electron Agr 205:107636.

Guo S, Yoon SC, Li L, Wang W, Zhuang H, Wei C, et al., 2023. Recognition and positioning of fresh tea buds using YOLOv4-lighted and ICBAM model and RGB-D sensing. Agriculture 13:518.

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition. Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas; pp. 770-778.

Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al., 2019. Searching for MobileNetV3. Proc. IEEE/CVF International Conference on Computer Vision, Seoul; pp. 1314-1324.

Hu J, Shen L, Sun G, 2018. Squeeze-and-excitation networks. Proc. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City; pp. 7132-7141.

Huang H, Chen Z, Zou Y, Lu M, Chen C, Song Y, et al., 2024. Channel prior convolutional

attention for medical image segmentation. Comput Biol Med 178:108784.

Kumar PVS, Basheer S, Ravi R, Thakur MS, 2011. Comparative assessment of tea quality by various analytical and sensory methods with emphasis on tea. J Food Sci Technol 48:440-446.

Liu P, Qiu X, Huang X, 2017. Adversarial multi-task learning for text classification. arXiv:1704.05742

Liu Y, Shao Z, Hoffmann N, 2021. Global attention mechanism: retain information to enhance channel-spatial interactions. arXiv:2112.05561.

Ma N, Zhang X, Zheng HT, Sun J, 2018. ShuffleNet V2: practical guidelines for efficient CNN architecture design. Proc. European Conference on Computer Vision, Munich; pp. 116-131.

Ruder S, Plank B, 2018. Strong baselines for neural semi-supervised learning under domain shift. arXiv:1804.09530

Salzmann M, Ek CH, Urtasun R, Darrell T, 2010. Factorized orthogonal latent spaces. Proc. 13th Int. Conf. on Artificial Intelligence and Statistics, Chia; pp. 701-708.

Song Y, Morency LP, Davis R, 2012. Multi-view latent variable discriminative models for action recognition. Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence; pp. 2120-2127.

Wang S, Wu D, Zheng X, 2023a. TBC-YOLOv7: a refined YOLOv7-based algorithm for tea bud grading detection. Front Plant Sci 14:1223410.

Wang T, Zhang K, Zhang W, Wang R, Wan S, Rao Y, et al., 2023b. Tea picking point detection and location based on Mask-RCNN. Inf Process Agric 10:267-275.

Woo S, Park J, Lee JY, Kweon IS, 2018. CBAM: convolutional block attention module. Proc. European Conference on Computer Vision, Munich; pp. 3-19.

Xie S, Sun H, 2023. Tea-YOLOv8s: a tea bud detection model based on deep learning and computer vision. Sensors (Basel) 23:6576.

Yang M, Yuan W, Xu G, 2024. YOLOX target detection model can identify and classify several types of tea buds with similar characteristics. Sci Rep 14:2855.

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J, 2024. DETRs beat YOLOs on real-time object detection. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle; pp. 16965–16974.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, 2016. Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas; pp. 2921-2929.



Figure 1. Misdetection sample of tea stems as buds.



Figure 2. The image samples of the ZC108 dataset.



Figure 3. The partial instances of tea stems **(a)** and tea buds **(b)**.

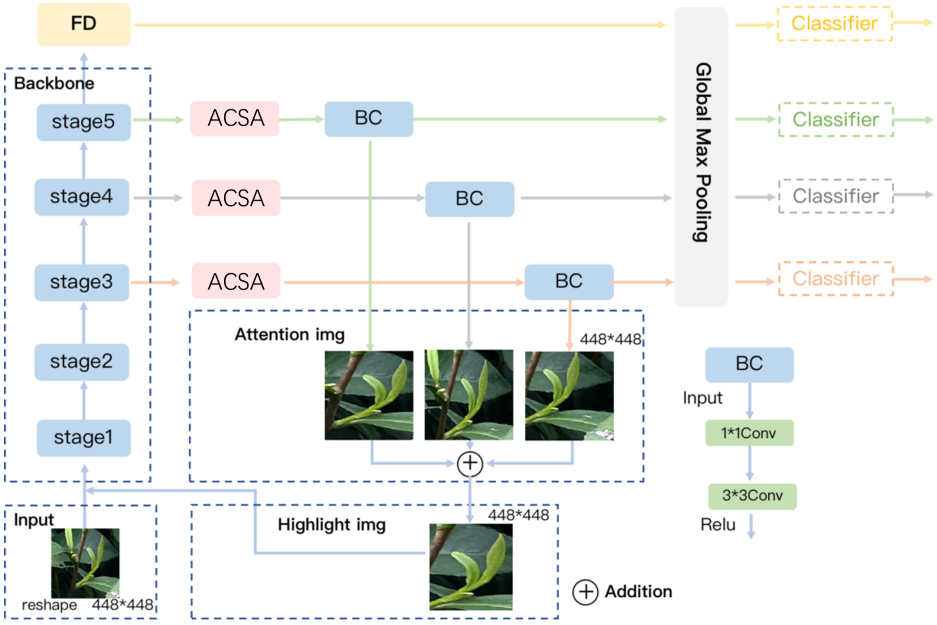


Figure 4. The overall architecture of MSIFD-Net. BC is the basic convolution module; FD is the feature decoupling module; ACSA is the adaptive channel-spatial attention module.

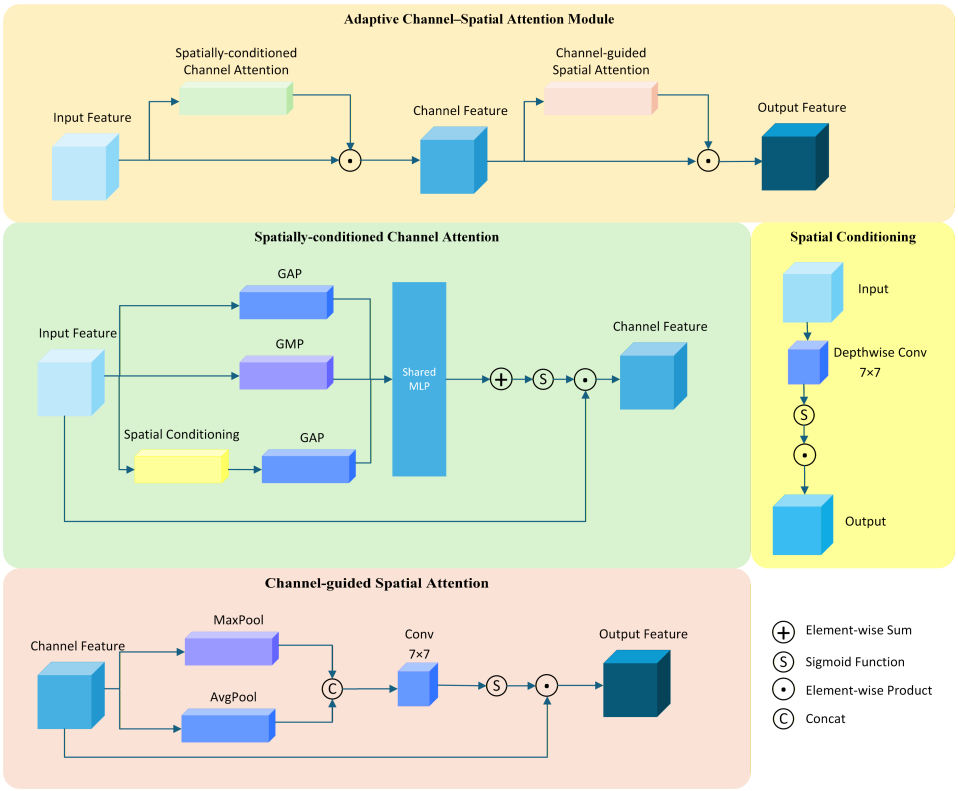


Figure 5 . Detailed structure of the proposed Adaptive Channel–Spatial Attention (ACSA) module. It adaptively refines features through bidirectional interaction: the top block performs channel attention conditioned on spatial context, while the bottom block executes spatial localization guided by channel semantics.

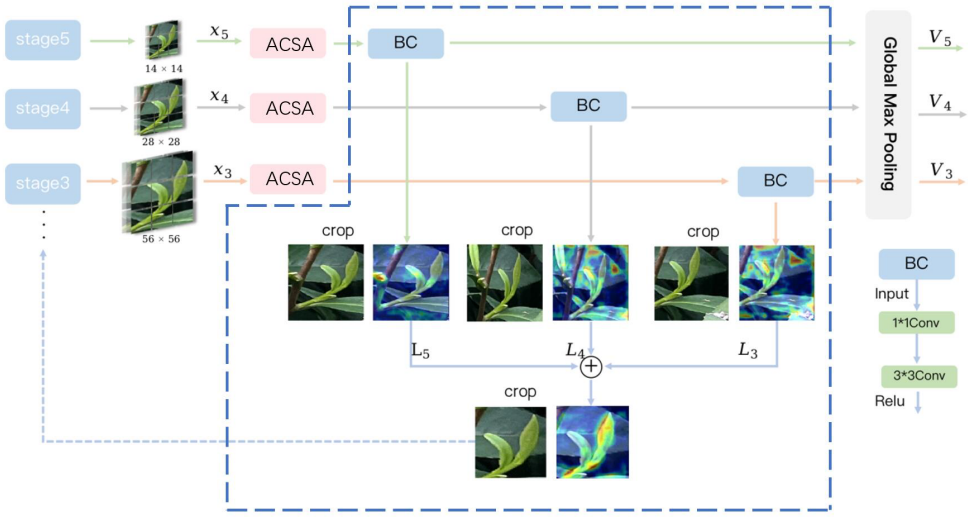


Figure 6. Schematic diagram of the multi-scale interaction structure.

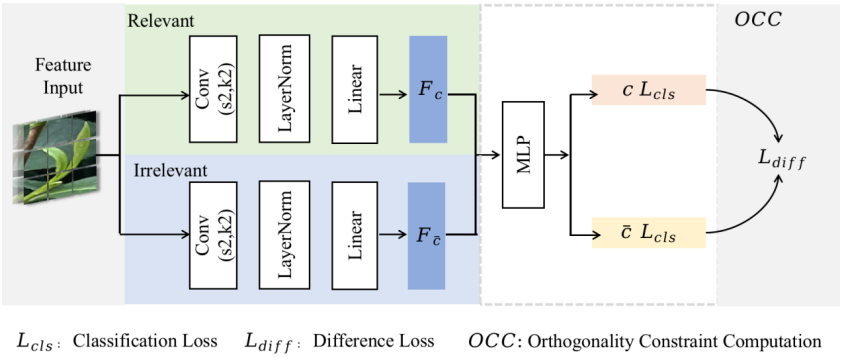


Figure 7. Feature decoupling module structure.

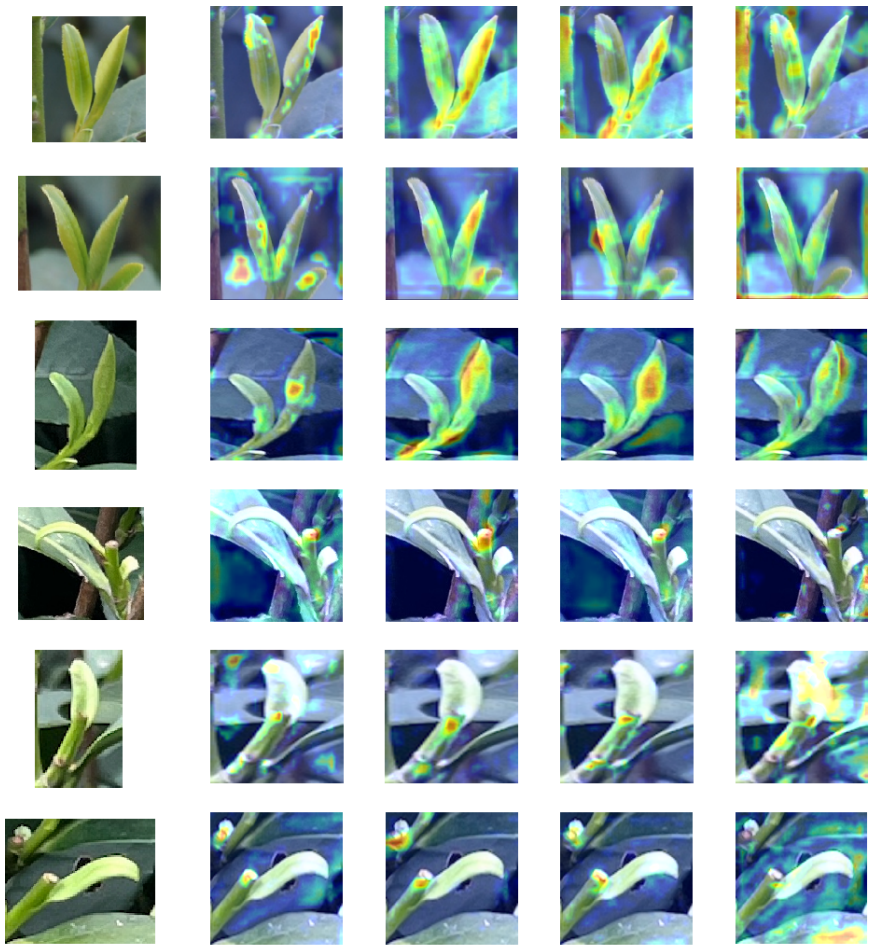


Figure 8. Visualization heatmaps of the impact of different scale interaction structures on network performance: two-scale (a), three-scale (b), four-scale (c), and five-scale (d) networks.



Figure 9. The recognition result of MSIFD-Net **(a)** shows the prediction results of YOLOv11, **(b)** shows the filtering results of MSIFD-Net.

Table 1. The specific data split of the ZC108 dataset.

Category	Train	Val	Test	Total

Bud	5760	1845	2065	9670
Stem	1920	491	573	2984

Table 2. Performance of MSIFD-Net using different backbone networks.

Backbone	Params(M)	FLOPs(G)	ACC(%)	P(%)	F1
ResNet18	11.24	11.36	92.10	94.32	0.921
MobileNet V3 Large	5.48	1.44	91.24	93.45	0.912
MobileNet V3 Small	2.54	0.38	89.18	89.47	0.895
ShuffleNet V2 0.5×	0.37	0.27	88.38	87.81	0.873
ShuffleNet V2 1.0×	1.29	0.95	93.08	91.90	0.905
ShuffleNet V2 1.5×	2.52	1.92	93.25	93.53	0.927
ShuffleNet V2 2.0×	5.42	3.76	**93.85**	**95.12**	**0.933**

Table 3. Performance of MSIFD-Net using different attention mechanisms.

Attention	Params(M)	ACC (%)	P (%)	F1(%)
—	5.42	93.85	94.90	0.921
SE	5.82	94.17	95.30	0.925
CBAM	5.85	95.13	96.60	0.936
GAM	5.95	94.79	96.10	0.933
CPCA	6.13	94.10	95.20	0.924
ACSA	**5.83**	**96.68**	**98.40**	**0.961**

Table 4. Performance of network structures at different scales.

#	Variants	ACC (%)	P (%)	F1
(a)	2-scale classification	95.76	95.97	0.946
(b)	3-scale classification	**96.38**	**96.65**	**0.950**
(c)	4-scale classification	94.75	95.13	0.933
(d)	5-scale classification	94.11	94.83	0.929

Table 5. Results of ablation experiments.

Backbone	FD	ACSA	MSI	ACC(%)	P (%)	F1
ShuffleNet V2	-	-	-	93.85	94.90	0.921
	√	-	-	94.62	96.05	0.941
	√	√	-	97.12	98.26	0.962
	√	√	√	**97.35**	**98.92**	**0.975**

Table 6. Comparison of performance for tea buds before and after adding MSIFD-Net on the ZC108 dataset.

Model	AP (%)	FDR(%)	FPS	Model	AP (%)	FDR(%)	FPS
YOLOX	74.32	26.85	110		84.96	1.51	92
YOLOv8	76.15	25.21	140		87.12	1.33	113
RT-DETR	77.41	24.58	130	+MSIFD-	86.85	1.42	105
YOLOv9	78.56	23.64	120	Net	89.34	1.31	100
YOLOv10	79.12	23.91	180		88.76	1.46	137
YOLOv11	81.05	22.22	195		91.28	1.10	154