Computers and Electronics in Agriculture 237 (2025) 110481

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

TDDet: A novel lightweight and efficient tea disease detector

Yange Sun [a,b], Zhihao Li [a], Huaping Guo [a,b,*], Yan Feng [a,*], Yongqiang Tang [c], Wensheng Zhang [c], Jingqiu Gu [d]

[a] School of Computer and Information Technology, Xinyang Normal University, Xinyang, 464000, China
[b] Henan Key Laboratory of Tea Plant Biology, Xinyang Normal University, Xinyang, 464000, China
[c] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation Chinese Academy of Sciences, Beijing, 100190, China
[d] National Engineering Research Center for Information Technology in Agriculture, Beijing, 100097, China

ARTICLE INFO

Dataset link: https://github.com/hpguo1982/TDDet

Keywords:
Tea disease
Object detection
Partial convolution
Efficient multiscale attention

ABSTRACT

Tea diseases cause significant economic losses to the tea industry every year, and thus developing a rapid and accurate tea disease detector is of great significance for assisting farmers in preventing diseases and increasing their income. Therefore, this paper proposes a lightweight and efficient detector called TDDet to quickly and accurately detect tea diseases. TDDet is mainly composed of two key innovations: feature extraction and feature aggregation. For feature extraction, we use lightweight depthwise separable convolution to reduce the computational load and enhance the ability to extract key local features in images of tea diseases. In addition, attention mechanisms including channel-, spatial-, and self-attentions, are employed to enable the model to focus on the most important parts of tea diseases, thereby improving the performance of the model. For feature aggregation, we propose a novel Cross-scale Feature Fusion (CFF) module to focus on tea disease areas, boosting the model's sensitivity to feature details. Based on CFF, TDDet repeatedly fuses multiscale features of different levels in a top-down and bottom-up manner, enhancing feature representation capability. Besides, a lightweight and efficient upsampling module, called Dysample, is used to reduce computational costs and improve model performance by dynamically adjusting the sampling rate of feature maps. Experimental results demonstrate that TDDet with fewer parameters outperforms other state-of-the-art object detection models, enabling fast and accurate identification of tea diseases. Our code and dataset are available at https://github.com/hpguo1982/TDDet.

1. Introduction

Tea is an important economic crop, capable of enhancing the human immune system and helping to prevent various diseases (Bag et al., 2022; Xia et al., 2020; Xu et al., 2021; Wei et al., 2024). The yield and quality of tea are key factors of tea farmers' income, making them vital for sustainable livelihoods. However, tea cultivation encounters significant challenges from various diseases that adversely affect both yield and quality, leading to considerable economic losses (Liu et al., 2020; Krishnakumar et al., 2024). Therefore, timely and accurate detection is essential to reduce potential losses and ensure effective control measures.

Traditional detection methods, which often rely on visual inspections by experienced personnel, are not only time-consuming but also susceptible to human error. Recent field studies reveal that traditional methods suffer from a 5–7 day detection delay for early-stage anthracnose lesions, with an accuracy below 75% in complex field conditions (Zhang et al., 2024; Ye et al., 2024a). Recently, deep learning, particularly Convolutional Neural Networks (CNNs) (Gu et al., 2018), has become a cornerstone of modern object detection and significantly enhanced tea disease detection accuracy (Krizhevsky et al., 2017; Szegedy et al., 2015; Liu et al., 2016; Sun et al., 2022; Liu et al., 2021b; Xue et al., 2022; Ashok et al., 2020; Xiong et al., 2024). For example, YOLOv8-RMDA achieves an mAP of 89.76% in identifying presymptomatic infections (24–72 h post-pathogen invasion) (Ye et al., 2024b). Compared with traditional methods, YOLOv8-RMDA's small object detection head significantly reduces the missed detection rate of lesions smaller than 32×32 pixels.

In the tea disease detection task, the success of CNN-based detection methods mainly relies on two key aspects: (a) extracting features with robust representation from the backbone network and (b) strengthening feature aggregation of the neck network.

(a) **Feature Extraction**: Feature extraction is the process of automatically constructing new features from raw data, transforming the original data into a set of features that are more representative and useful for a specific task. In tea disease detection,

* Corresponding authors.
E-mail addresses: hpguo@xynu.edu.cn (H. Guo), yfeng@xynu.edu.cn (Y. Feng).

https://doi.org/10.1016/j.compag.2025.110481
Received 21 October 2024; Received in revised form 12 April 2025; Accepted 29 April 2025
Available online 16 May 2025

feature extraction typically involves identifying and highlighting important features, such as edges, textures, and shapes. For example, Li et al. (2022) proposed an improved Mask R-CNN (Hossain et al., 2018), using the ResNet (He et al., 2016) as the feature extraction network to improve the detection precision of diseases with different shapes. Hu et al. (2021) integrated Feature Pyramid Networks (FPN) (Lin et al., 2017a) into Faster R-CNN (Ren et al., 2016) to enhance recognition of blurred, occluded, and small diseased leaves. Xue et al. (2023) developed YOLO-Tea, a tea disease detection model incorporating the Global Context Network (GCNet) (Cao et al., 2020), improving the model's focus on diseased regions. Bhuyan et al. (2024) introduced a Res4Net-based model with a Convolutional Block Attention Module(CBAM) (Woo et al., 2018), improving the model's feature extraction for different diseases.

(b) **Feature Aggregation**: Feature aggregation is a technique to combine features from different levels to enhance feature expression, leading to improved performance of the model. For example, Bao et al. (2022) integrated a multiscale feature aggregation module into RetinaNet (Lin et al., 2017b), enabling hierarchical feature aggregation that enhances semantic representation and improves detection precision. Li and Zhao (2025) proposed the ECA-ResNet50, which employs a multilayer small-kernel convolution strategy to improve cross-scale feature aggregation and mitigate the interference of complex backgrounds. Xia et al. (2024) improved YOLOv7 by adopting the lightweight MobileNeXt backbone to reduce computation and enhance efficiency. They also introduced a two-layer routing attention mechanism to enhance feature aggregation, improving the model's ability to capture disease details and textures.

Attention mechanisms, including channel-based attention, spatial-based attention, and self-attention, are often employed to enhance feature representation due to their ability to dynamically focus on the most relevant parts of the input data (Liu et al., 2023a; Bala et al., 2024). Channel-based attention dynamically adjusts feature weights at the channel level, while spatial-based attention highlights specific regions within an image. These two mechanisms are often combined to enhance the performance of convolutional neural networks for tea disease analysis. For instance, Liang et al. (2025) proposed LTDDN, a lightweight model for tea disease classification, leveraging the Channel Focus Attention (CFA) mechanism to enhance disease feature representation while effectively suppressing background interference, thereby improving recognition accuracy. Lin et al. (2023) developed TSBA-YOLO, incorporating Shuffle Attention (Zhang and Yang, 2021) to improve small target detection. (Wang et al., 2023b) introduced Global Attention Mechanism (GAM) and Convolutional Block Attention Module (CBAM) into YOLOv5, enhancing fine-grained feature extraction for tea disease images. Moreover, self-attention mechanisms, a specialized form of spatial attention, have been utilized to capture global interactions between image regions. For example, Sun et al. (2023) proposed TeaDiseaseNet, a multiscale self-attention detection model, improving recognition of variable-scale disease symptoms. Yang et al. (2023) developed YOLOv7-Tiny for tea disease detection by integrating the Biformer (Zhu et al., 2023) and a dynamic attention mechanism to enhance feature representation and improve detection accuracy. He et al. (2024) proposed an enhanced YOLOv7 (Wang et al., 2023a) network for tea disease recognition, integrating the Vision Transformer's attention mechanism to refine feature extraction and improve detection accuracy.

Recent advancements in detection methods for tea diseases have shown significant improvements, as discussed above. However, achieving ideal results remains challenging in certain complex scenarios. As shown in Fig. 1, YOLOv8 (Jocher et al., 2023) often generates false positives by highlighting areas unrelated to actual disease symptoms. RT-DETR (Zhao et al., 2024) tends to miss subtle lesions, resulting in incomplete or imprecise detection. YOLOv10 (Wang et al., 2024a) struggles to capture fine disease details. In addition, most current deep learning methods for tea disease detection struggle with high computational complexity and slow inference speed, making them unsuitable for real-time processing of high-resolution images or large-scale data, which limits their practical application.

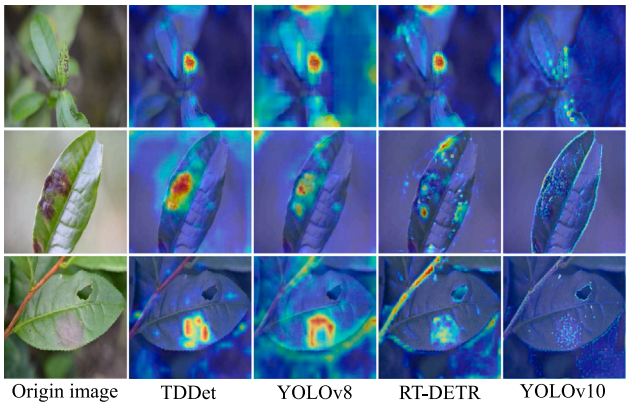

Fig. 1. Heatmaps generated by our method and three object detection methods.

To tackle the challenges discussed above, we propose TDDet, a lightweight and efficient detector that enhances feature extraction and aggregation through attention mechanisms. For feature extraction, TDDet integrates Multi-Query Attention (MQA) and Depthwise Separable Convolution (DW) to optimize computational efficiency while preserving rich feature details. MQA reduces computation and memory overhead by sharing queries and independently mapping keys and values, ensuring robust feature representation. Meanwhile, DW minimizes parameters and improves processing efficiency without compromising feature quality. For feature aggregation, we introduce a Cross-scale Feature Fusion (CFF) module, leveraging Efficient Multiscale Attention (EMA) and Partial Convolution (PConv). EMA explores spatial location relationships through large receptive fields, enhancing multiscale spatial information, while PConv selectively retains relevant input regions, reducing computational costs and improving aggregation efficiency. Additionally, DySample is employed for efficient and lightweight dynamic upsampling, further enhancing feature aggregation and detection performance.

In summary, the main contributions of this study are as follows:

- We propose a novel end-to-end TDDet for automatically detecting and recognizing tea diseases in natural scene images. We utilize MQA and DW to capture more detailed and informative features relevant to tea diseases, enhancing the extraction capabilities of features from diseased leaves.
- Based on EMA and PConv, we propose a CFF model that selectively retains portions of the input tensor for efficient feature fusion. This model reduces computational demands while improving multiscale feature aggregation.
- Experimental results demonstrate that TDDet achieves remarkable detection results, with a mAP of 94.33%, a Precision of 91.33%, and a Recall of 92.94%, outperforming state-of-the-art object detection models.

The remainder of the paper is organized as follows: Section 2 introduces our TDDet, Section 3 presents the dataset and presents the experimental results, and finally, the work is summarized in Section 4.

2. Methods

2.1. Overall architecture

The proposed TDDet is a lightweight model capable of operating on mobile or edge devices with constrained resources. Following the

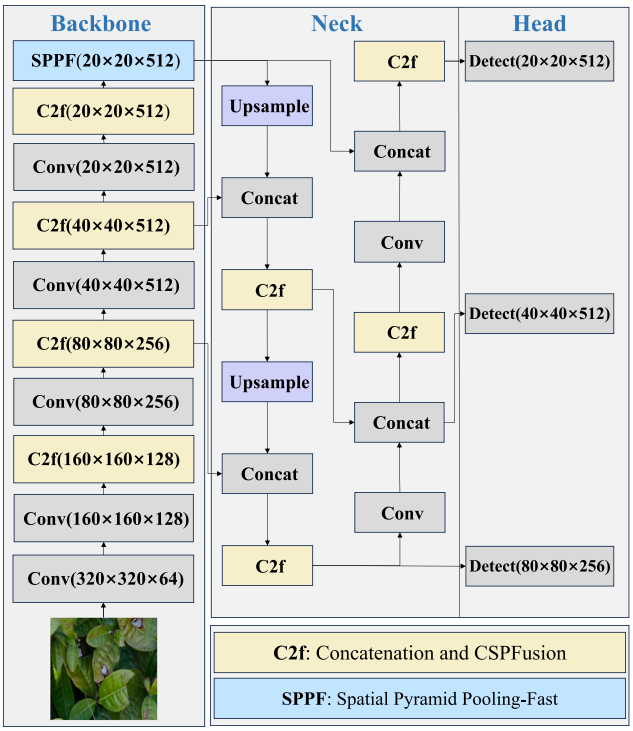

Fig. 2. Diagram of the YOLOv8 module structure.

YOLOv8 architecture (Jocher et al., 2023), as shown in Fig. 2, TDDet retains its structure but introduces several architectural enhancements to optimize feature extraction and aggregation. Fig. 3 shows the architecture of our TDDet, which consists of three key modules: the backbone, the neck, and the head.

To balance feature extraction and computational efficiency, the backbone incorporates Depthwise Separable Convolution (DW) to reduce complexity and the Universal Inverted Bottleneck (UIB) to enhance multi-level feature extraction capabilities. The neck employs bidirectional feature pyramid aggregation to enhance multiscale feature interaction, where the top-down path uses DySample-based upsampling to refine spatial details and the bottom-up path employs convolution-based downsampling to enrich semantics. In addition, the Cross-scale Feature Fusion (CFF) module is used by both paths to maximize feature complementarity and information flow. Finally, TDDet uses the multiscale features produced by CFF as inputs to its head module, enabling accurate and robust prediction.

2.1.1. Backbone network

We use MobileNetv4 (Qin et al., 2024) as the backbone for feature extraction, adopting a hierarchical modular to enhance feature extraction capabilities, as shown in Fig. 3(a). MobileNetv4 uses five stages for feature extraction. Following the convolutional operations of the first stage, the UIB modules are employed in the second and third stages. The UIB uses various combinations of depthwise and pointwise convolutions to create depthwise separable convolution blocks to enhance feature extraction efficiency. In the fourth and fifth stages, we employ the UIB and MQA modules to further enhance feature extraction by capturing complex dependencies. MQA is an improvement over the traditional multi-head self-attention, which uses shared keys and values to simplify and optimize attention computation efficiency. In addition, we employ the Spatial Pyramid Pooling-Fast (SPPF) module (Jocher, 2020) to enhance multiscale features through multi-level pooling operations. The details of the UIB and the MQA are introduced in Sections 2.2 and 2.3, respectively.

2.1.2. Neck network

The neck network employs a bidirectional feature pyramid aggregation to refine and fuse the outputs from the backbone network, as illustrated in Fig. 3(b). The bidirectional pathways, encompassing both top-down and bottom-up flows, enhance the information exchange and integration across features of varying scales. For the top-down path, the high-level features (with semantic information) are upsampled by the enhanced DySample and then fused with low-level features (with spatial information) using concatenate operation, followed by the proposed CFF to output a refined feature representation. For the bottom-up path, we employ a strategy similar to the top-down path with the exception that we reduce the dimensions of the feature map rather than expanding them. Specifically, low-level features are downsampled using convolutional operation with a stride step equal to 2 and then fused with high-level features using CFF to output a complementary feature representation that enhances the information flow between different scales. The details of the CFF and the Dysample are introduced in Sections 2.4 and 2.5.

2.1.3. Head

The head receives the multiscale features produced by the CFF modules in the neck network, which are subsequently processed through three detection branches to enable accurate localization and classification of disease regions across various scales.

2.2. Multi-Query Attention (MQA)

We implement MQA within our tea disease detection model to reduce the training burden while enhancing performance, as shown in Fig. 4. MQA is an innovative attention mechanism that shares keys and values across all heads while independently processing queries. This design allows the model to efficiently focus on various aspects of the input, such as distinct features of tea leaves affected by different diseases, thereby preserving the richness of self-attention mechanisms. By sharing keys and values across attention heads, MQA reduces memory bandwidth requirements, which is crucial for handling high-resolution images. This shared approach leads to a substantial reduction in computational complexity. Specifically, traditional Multi-Head Attention has a complexity of $O(B \cdot H \cdot (N \cdot M + N^2 + M^2))$, where B is the batch size, H is the number of heads, N is the input sequence length, and M is the context sequence length. MQA simplifies this complexity to $O(B \cdot (N \cdot M + N^2))$.

2.3. Universal Inverted Bottleneck (UIB)

We integrate the UIB module into our TDDet to enhance the efficiency and effectiveness of information processing. Fig. 5(left) shows the structures of the four variations of the UIB. The UIB is an innovative component that leverages Depthwise Separable Convolutional Blocks (DW) (Howard et al., 2017). DW is a type of convolution operation that separates the mixing of input channels (Depthwise Convolution) from the spatial correlation (Pointwise Convolution), as shown in Fig. 5(right), significantly reducing convolutional parameters and computational costs.

In the backbone architecture Fig. 3(a), each stage employs a tailored UIB variant to enhance performance: Stage 2 adopts variant (d), Stage 3 uses variant (b), Stage 4 applies variant (c), and Stage 5 incorporates variant (a).

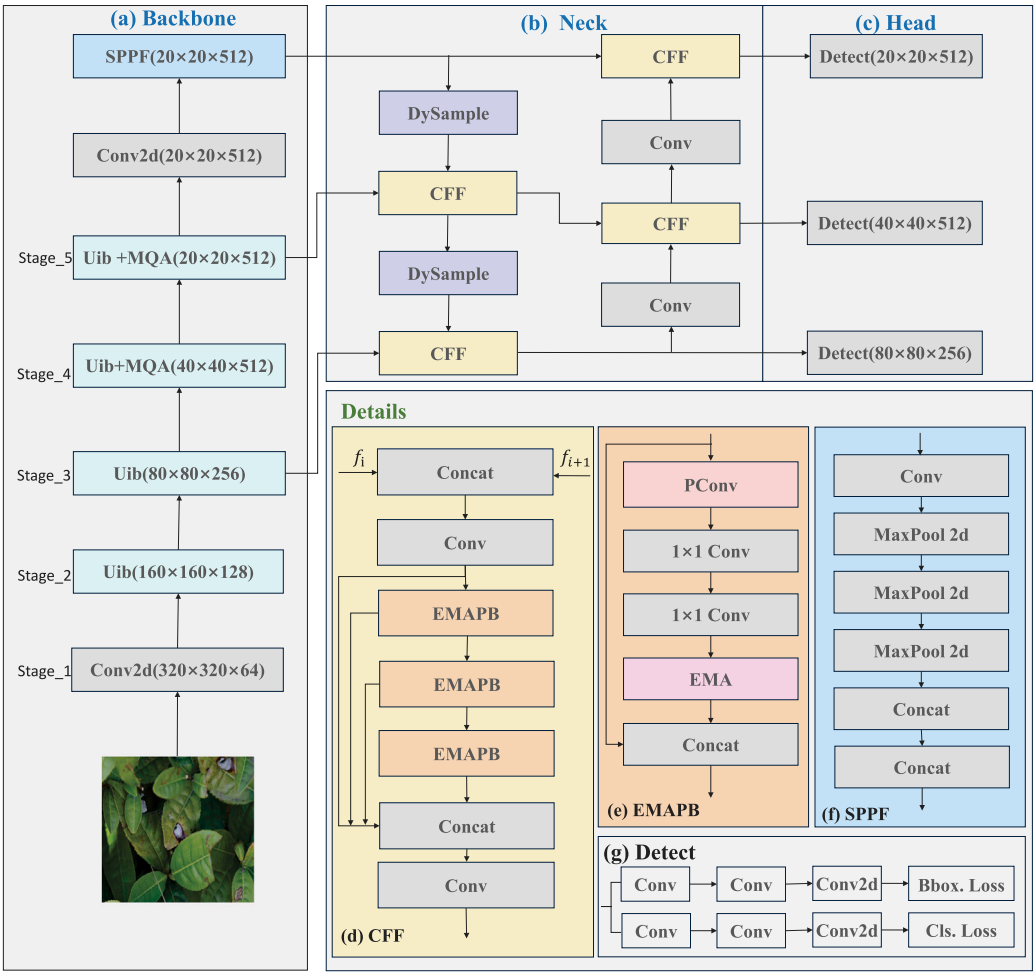

Fig. 3. Overall structure of the TDDet network model. (a) The Backbone extracts multi-scale features using UIB and MQA modules. (b) The Neck enhances feature fusion through DySample and CFF modules. (c) The Head detects objects at different scales using multi-resolution feature maps. (d) The CFF module performs feature aggregation with EMAPB blocks. (e) The EMAPB module refines features with PConv and EMA mechanisms. (f) The SPPF module compresses features while retaining spatial information. (g) The Detect module predicts object locations and classifications.

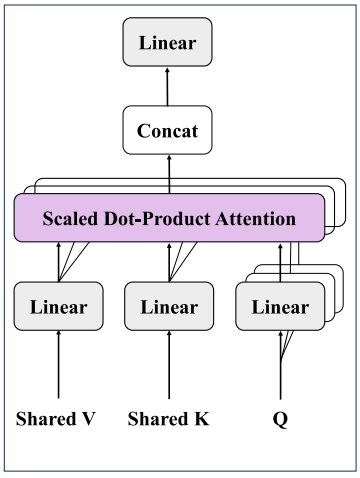

Fig. 4. Diagram of the Multi-Query Attention module structure.

2.4. Cross-scale Feature Fusion (CFF)

The proposed CFF module is employed to enhance feature aggregation within the neck network, as shown in Fig. 3(b). Fig. 3(d) shows the structure of CFF.

The CFF takes the low-level feature f_i and the high-level feature f_{i+1} as inputs and concatenates them as f_i^{concat}, followed by a convolutional block that includes convolution, batch normalization, and ReLU activation. Formally,

$$f_i^{conv} = \mathrm{ReLU}\left(\mathrm{BatchNorm}\left(\mathrm{Conv}\left(f_i^{concat}\right)\right)\right) \tag{1}$$

f_i^{conv} is then split into two subfeatures $f_i^{sub_1}$ and $f_i^{sub_2}$. $f_i^{sub_2}$ is fed into three successive Efficient Multiscale Attention Projection Blocks (EMAPB), which are essential for refining the feature representation relevant to various tea disease symptoms. Formally,

$$\begin{aligned} f_i^1 &= \mathrm{EMPAB}\left(f_i^{sub_2}\right) \\ f_i^2 &= \mathrm{EMPAB}\left(f_i^1\right) \\ f_i^3 &= \mathrm{EMPAB}\left(f_i^2\right) \end{aligned} \tag{2}$$

f_i^1, f_i^2 and f_i^3 are subsequently fused and processed by a final convolution to generate the output feature map f_i^E. Formally,

$$f_i^E = \mathrm{Conv}\left(\mathrm{Concate}\left(f_i^{sub_1}, f_i^1, f_i^2, f_i^3\right)\right) \tag{3}$$

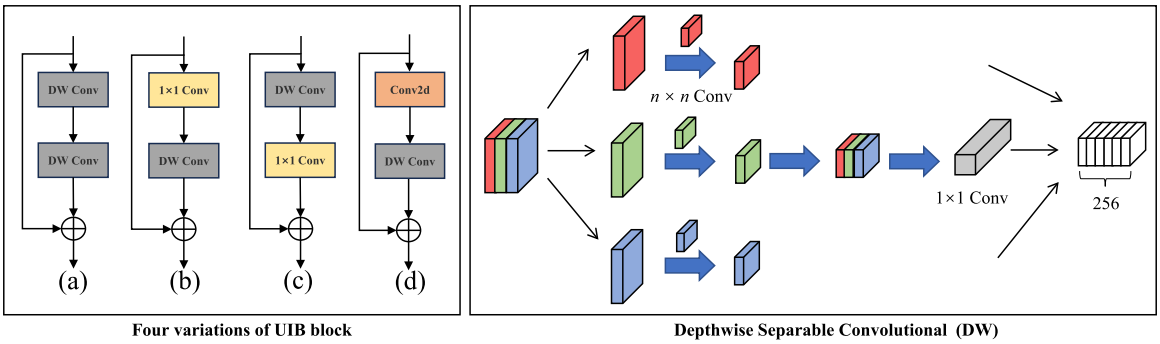

Fig. 5. Diagram of the UIB and DW Conv module structure.

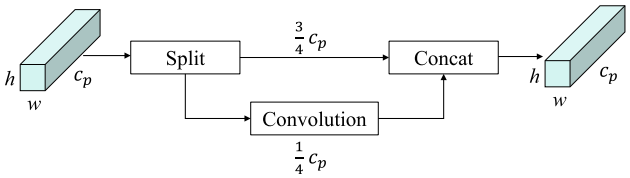

Fig. 6. Diagram of the Partial Convolution module structure.

EMAPB plays a vital role in the CFF module by progressively enhancing feature representation, and Fig. 3(e) shows the details of EMAPB. Let f denote the input of EMAPB. f undergoes a partial convolution (PConv) operation, followed by two consecutive 1×1 convolutional layers, to extract spatial features f_s as follow:

$$f_s = \mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{PConv}\left(f\right)\right)\right) \tag{4}$$

Subsequently, f_s enters an Efficient Multiscale Attention (EMA) module, which aggregates multiscale information and establishes dependencies to enhance the global context capture capability. The output of the EMA module is then fused with the initial feature maps for a shortcut connection. Formally,

$$f_{st} = f + \mathrm{EMA}\left(f_s\right) \tag{5}$$

The details of PConv and EMA are presented in Sections 2.4.1 and 2.4.2.

2.4.1. Partial Convolution (PConv)

Fig. 6 shows the structure of PConv. We split the input f into two parts: $\frac{3}{4}$ of the f remain unchanged, while the remaining undergo a conventional convolution to extract spatial features, and then concatenate the two parts. In this way, we optimize computational efficiency while maintaining overall feature integrity. Generally, we consider the input and output feature to have the same number of channels. The floating-point operations (FLOPs) of a PConv is $h \times w \times k^2 \times c_p^2$ where h and w denote the height and width of the feature map, k is the convolution kernel size, and c_p represents the number of channels involved in the standard convolution. Here, c_p is one-fourth of the original channel, and thus the FLOP for PConv is one-sixteenth of a regular convolution.

2.4.2. Efficient Multiscale Attention (EMA)

We introduce an EMA module to enhance feature representation while minimizing computational overhead. EMA employs strategies including feature grouping, parallel subnetworks, and cross-space learning to effectively learn channel descriptions without reducing channel dimensionality, thereby generating improved pixel-level attention.

The structure of EMA is illustrated in Fig. 7. The input feature $f \in \mathbb{R}^{c\times h\times w}$ is divided into g sub-groups, i.e., $\left\{ f_i \in \mathbb{R}^{\frac{c}{g}\times h\times w} \mid i = 0, 1, \ldots, g-1 \right\}$, with $f = \bigcup_{i=0}^{i=g-1} f_i$ and $f_i \cap f_j = \emptyset$ for $i \neq j$. EMA utilizes three parallel paths to extract attention-weight descriptors for each f_i: two paths route through the 1×1 branch, while one uses the 3×3 branch. The two 1×1 branches process features similarly to CA (Hou et al., 2021), each encoding the channel along a spatial direction to obtain attentional features f_{att}. The 3×3 branch employs convolution with $\frac{c}{g}$ channels and a kernel size of 3×3 to capture feature representation f_{pre}. Then f_{att} and f_{pre} are combined through cross-spatial learning, enhancing feature aggregation across different spatial dimensions. This allows EMA to effectively capture both global and local spatial dependencies, which is crucial for detecting tea diseases where tea disease symptoms may be closely located and have similar shapes. Such proximity can lead to missed or false detections, emphasizing the need for precise spatial structure preservation within the channels. Consequently, EMA not only encodes cross-channel information to adjust the importance of different channels, but also maintains detailed spatial structure information.

2.5. DySample module

Upsampling is a commonly used operation in object detection tasks. However, traditional upsampling techniques like bilinear interpolation suffer from pixel distortion and computational inefficiencies (Wang et al., 2019). To address these challenges, we integrate a dynamic upsampler called DySample into our TDDet. DySample dynamically upsamples points by combining point sampling techniques with a trainable sampling strategy. This method eliminates the computationally expensive dynamic convolutional layers and additional sub-networks, thereby reducing the computational burden and enhancing the model's efficiency. The structure of the DySample is shown in Fig. 8. Specifically, DySample operates by first applying a linear layer to the input feature f to generate an initial offset volume, then using the pixel shuffle (Shi et al., 2016) operation to get the offset O. The offset O is subsequently added to the original sampling grid G, generating the sampling set S. The grid_sample function (Gyapong and Remme, 2001) is finally employed to resample the sampling set S, yielding the upsampled feature map f'. By sidestepping the intricacies of dynamic convolution, this method achieves high efficiency in terms of low latency and memory usage.

2.6. Loss function

In this study, we adopt the Enhanced Intersection over Union (EIoU) (Zhang et al., 2022) as our loss function, defined as

$$L_{\mathrm{EIoU}} = L_{\mathrm{IoU}} + L_{\mathrm{dls}} + L_{\mathrm{arl}} \tag{6}$$

where

$$\begin{aligned} L_{\mathrm{IoU}} &= 1 - IoU \\ L_{\mathrm{dls}} &= L_{\mathrm{IoU}} + \frac{\rho^2\left(b, b^{gt}\right)}{\left(w^c\right)^2} \\ L_{\mathrm{arl}} &= -L_{\mathrm{IoU}} + \frac{\rho^2\left(w, w^{gt}\right)}{\left(w^c\right)^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(h^c\right)^2} \end{aligned} \tag{7}$$

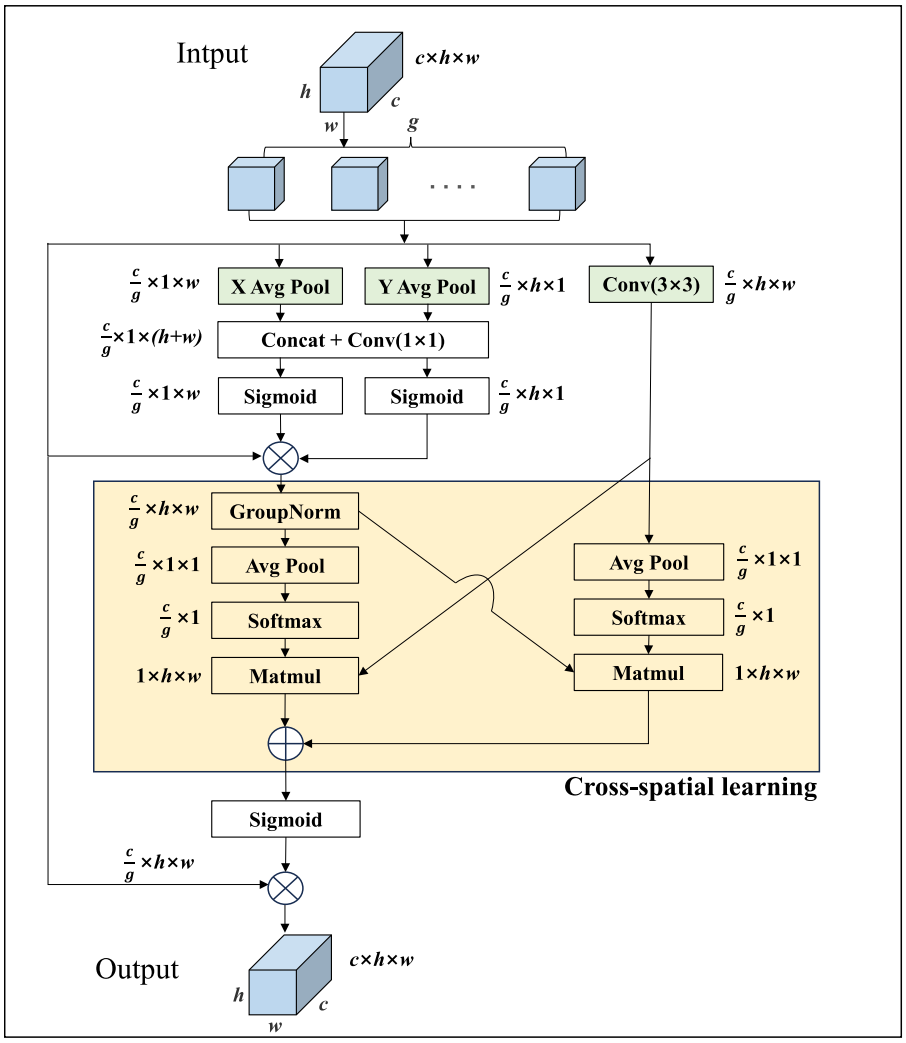

Fig. 7. Diagram of the Efficient Multiscale Attention module structure.

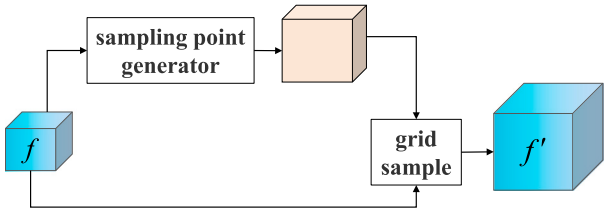

Fig. 8. Diagram of the Dysample module structure.

where IoU (Yu et al., 2016) is a measure of the overlap between the predicted and the true boxes. L_{dls} is the Distance Loss (DL) (Yang et al., 2021), which quantifies the distance between the centers of the predicted and ground truth boxes, normalized by the candidate box width. This loss helps reduce center offsets and improve localization accuracy, which is crucial for detecting tea diseases, as the symptoms often appear as small and densely packed targets. L_{arl} is the Aspect Ratio Loss (ARL) (Bauters et al., 2011), measuring the differences in width and height between the predicted and ground truth boxes, normalized by the dimensions of the candidate box. The ARL aligns the predicted box shape with the target, enhancing detection accuracy for varying aspect ratios—crucial for identifying diverse symptoms of tea diseases that may exhibit similar shapes.

EIoU is suitable for tea disease detection as it effectively addresses multiple challenges, including small target detection via IoU, center alignment via DL, and densely clustered distributions via ARL. In contrast, alternative IoU variants exhibit notable limitations: GIoU (Rezatofighi et al., 2019) extends traditional IoU by incorporating non-overlapping regions but does not directly optimize localization precision. DIoU (Zheng et al., 2020) introduces center distance minimization but neglects aspect ratio consistency. CIoU (Zheng et al., 2020) balances these factors by considering aspect ratio constraints, but its optimization relies on predefined weighting factors.

3. Experiments

3.1. Datasets

3.1.1. Data collection

The tea disease dataset was collected by researchers from the Institute of Agricultural Economics and Information, Anhui Academy of Agricultural Sciences (Sun et al., 2019, 2023). Images were captured at a local tea plantation in Anhui Province using a Canon EOS 5D Mark IV DSLR camera equipped with a 50 mm macro lens. This high-resolution setup, featuring a 30.4-megapixel full-frame sensor, ensured precise capture of disease symptoms, including subtle variations in leaf texture, color changes, and lesion patterns. The macro lens was selected for its consistent focal length, minimal distortion, and precise close-up imaging, which is critical for fine-grained disease identification.

Table 1
Distribution of labeled samples per disease class.

Disease type	Number of images
TB	137
TRS	127
TCLB	243
TC	116
TRR	125
TALS	128
Total	**876**

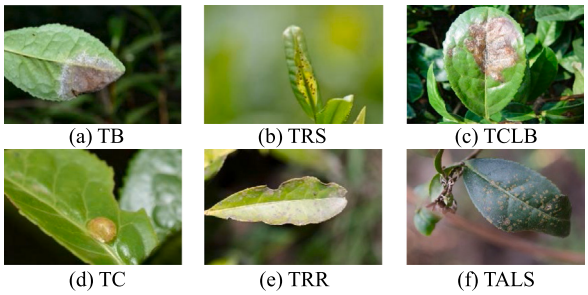

Fig. 9. Representative samples from the Tea dataset.

To enhance data reliability, three experienced tea plant pathologists validated the images, ensuring accurate identification and annotation of six distinct tea diseases: Tea Blight (TB), Tea red scab (TRS), Tea Cloud Leaf Blight (TCLB), Tea Cake (TC), Tea Red Rust (TRR), and Tea Algae Leaf Spot (TALS). Fig. 9 displays six images of tea leaves affected by various diseases, highlighting the diversity of symptoms captured in the dataset. The final dataset consists of 876 high-quality images, each with a resolution of 906 × 600 pixels, and the disease distribution is detailed in Table 1.

3.1.2. Data labeling

The original sample data lacked bounding box annotations for tea diseases, presenting a challenge for accurate disease detection. To address this issue, we utilized LabelImg (Gallian, 2012) to manually annotate bounding boxes for each image. This process was conducted in collaboration with three experienced tea tree pathologists to ensure annotation accuracy and consistency. As a result, a standardized dataset was established, providing a solid foundation for training and evaluating tea disease detection models.

3.1.3. Data preprocessing

One of the key challenges in tea disease detection is the limited availability of diverse and representative images, which often results in model overfitting and compromised generalization performance. To address this issue, we employed data augmentation techniques (Zhong et al., 2020), including brightness adjustment, hue adjustment, random rotation, noise addition, and mosaic augmentation, to increase the number of training samples, thereby improving the model's generalization ability, and preventing overfitting. Hue adjustment modifies the color tones of images to simulate varying lighting conditions. Brightness adjustment alters the intensity of light in images to replicate different illumination levels. Random rotation changes the orientation of images to capture disease features from multiple angles. Noise addition introduces random pixel variations to reduce the model's reliance on clean data. Mosaic augmentation combines and rearranges segmented parts of multiple images to create composite samples. Fig. 10 illustrates the outcomes of applying these augmentation techniques to the original image. Note: For compatibility with the YOLOv8-based model, all images were resized to 640 × 640 pixels.

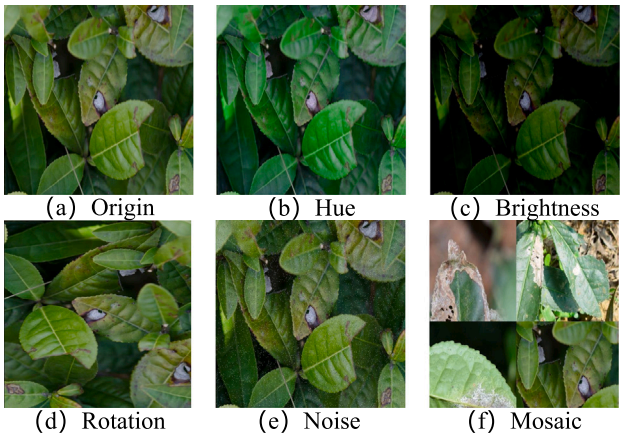

Fig. 10. Tea disease data enhancement strategies.

3.2. Experimental settings

3.2.1. Implementation details

Our experiments were conducted on an NVIDIA A100 GPU with 80 GB of memory, using a software environment of Python 3.9.1 and PyTorch 1.13.0 framework. Hyperparameters were primarily configured based on the default values of the YOLOv8 and the PyTorch framework, including an initial learning rate of 0.01, a momentum coefficient of 0.937, and an SGD optimizer with a weight decay of 0.0005. To balance computational efficiency and gradient stability, we selected a batch size of 16, ensuring robust feature learning while avoiding excessive memory consumption. The total number of epochs was set to 150, determined by observing convergence behavior, to achieve an optimal trade-off between underfitting and overfitting.

To further validate the robustness of our TDDet, we conducted a 5-fold cross-validation experiment: the dataset was randomly divided into five subsets, where one subset for testing and the other four subsets for training. This process was repeated five times, and the final results were averaged to obtain stable performance metrics. In addition, we used augmentation methods in Section 3.1.3 to augment the training dataset.

3.2.2. Evaluation metrics

The Mean Average Precision (mAP) is utilized as the principal metric for assessing the performance of our TDDet, defined as:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i. \tag{8}$$

where N is the total number of classes. AP_i is the Average Precision (AP) for the ith class. By averaging the AP across all classes, mAP provides a comprehensive evaluation of the model's effectiveness in detecting various types of tea diseases, ensuring a balanced assessment across different disease classes.

For each class, the AP is computed as the area under the precision–recall curve, which reflects the model's ability to distinguish diseased tea leaves from healthy ones under varying confidence thresholds.

$$\mathrm{AP} = \int_0^1 p(r)dr. \tag{9}$$

where $p(r)$ is the function representing the relationship between Precision(P) and Recall(R), respectively, defined as follows

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{10}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{11}$$

Table 2
Comparative experiments using 5-fold cross-validation for different models.

	Method	Total (%)	TALS (%)	TC (%)	TCLB (%)	TB (%)	TRR (%)	TRS (%)
Precision	SSD (Liu et al., 2016)	83.38 ± 3.65 •	85.18 ± 3.05 •	86.27 ± 2.17 •	82.86 ± 3.78 •	83.17 ± 3.35 •	87.48 ± 2.34 •	72.27 ± 1.86 •
	Faster R-CNN (Ren et al., 2016)	82.28 ± 3.77 •	84.75 ± 2.86 •	85.75 ± 2.34 •	82.07 ± 3.16 •	82.86 ± 2.85 •	88.11 ± 2.76 •	74.43 ± 1.91 •
	RetinaNet (Lin et al., 2017b)	85.47 ± 4.26 •	87.26 ± 3.76 •	87.71 ± 3.48 •	84.46 ± 4.15 •	85.36 ± 3.75 •	89.51 ± 3.18 •	73.72 ± 2.65 •
	YOLOv5 (Jocher, 2020)	89.65 ± 3.33 •	90.12 ± 3.17 •	91.43 ± 2.76 •	87.72 ± 3.27 •	88.43 ± 3.76 •	92.27 ± 2.17 •	75.75 ± 1.68 •
	YOLOv8 (Jocher et al., 2023)	90.52 ± 3.15 •	90.34 ± 2.49	92.19 ± 2.68 •	89.15 ± 3.07	90.18 ± 3.08 •	**94.71 ± 2.46**	81.36 ± 1.81 •
	DETR (Carion et al., 2020)	88.21 ± 3.32 •	89.73 ± 3.18 •	90.13 ± 2.44 •	88.37 ± 3.14 •	86.79 ± 3.17 •	91.73 ± 2.42 •	78.16 ± 2.08 •
	AX-RetinaNet (Bao et al., 2022)	88.16 ± 3.35 •	90.03 ± 3.72 •	90.13 ± 2.41 •	87.11 ± 3.01 •	86.79 ± 3.38 •	91.73 ± 2.75 •	78.16 ± 2.08 •
	RT-DETR (Zhao et al., 2024)	87.19 ± 2.41 •	90.18 ± 2.86	89.36 ± 2.37 •	87.68 ± 2.51 •	85.38 ± 2.18 •	90.67 ± 2.39 •	76.16 ± 1.76 •
	YOLOv9 (Wang et al., 2024b)	91.20 ± 2.53	90.19 ± 2.26	95.07 ± 1.58	**90.14 ± 2.19**	89.37 ± 1.83 •	94.57 ± 1.36	82.34 ± 1.87 •
	YOLOv10 (Wang et al., 2024a)	81.51 ± 2.56 •	83.78 ± 2.45 •	85.17 ± 2.34 •	81.65 ± 2.91 •	82.47 ± 3.09 •	85.75 ± 2.19 •	77.27 ± 1.86 •
	TDDet	**91.33 ± 2.14**	**90.87 ± 2.0**	**95.09 ± 1.27**	88.59 ± 2.08	**91.25 ± 1.97**	93.94 ± 0.95	**83.90 ± 1.37**
Recall	SSD (Liu et al., 2016)	81.10 ± 3.74 •	79.26 ± 3.18 •	80.09 ± 2.04 •	87.17 ± 3.46 •	84.17 ± 3.35 •	86.89 ± 2.16 •	69.71 ± 1.48 •
	Faster R-CNN (Ren et al., 2016)	85.62 ± 2.85 •	83.35 ± 2.18 •	84.79 ± 2.12 •	90.18 ± 2.83 •	87.61 ± 2.46 •	89.73 ± 2.18 •	73.46 ± 1.49 •
	RetinaNet (Lin et al., 2017b)	87.91 ± 4.17 •	85.42 ± 3.79 •	86.49 ± 3.16 •	92.49 ± 4.01 •	89.17 ± 3.43 •	93.17 ± 3.08 •	75.46 ± 1.58 •
	YOLOv5 (Jocher, 2020)	87.86 ± 3.46 •	86.43 ± 2.94 •	85.13 ± 2.23 •	92.68 ± 2.43 •	89.72 ± 3.38 •	92.73 ± 2.19 •	75.86 ± 1.76 •
	YOLOv8 (Jocher et al., 2023)	87.21 ± 3.24 •	85.17 ± 2.43 •	**90.47 ± 2.73**	92.46 ± 2.87 •	90.75 ± 2.78 •	91.97 ± 2.03 •	76.48 ± 1.31 •
	DETR (Carion et al., 2020)	86.55 ± 3.61 •	84.48 ± 3.48 •	85.48 ± 2.17 •	92.46 ± 2.89 •	89.68 ± 3.09 •	93.43 ± 2.17 •	77.16 ± 1.74 •
	AX-RetinaNet (Bao et al., 2022)	87.77 ± 3.17 •	84.36 ± 3.37 •	85.37 ± 2.41 •	93.19 ± 2.48 •	89.91 ± 3.17 •	92.39 ± 2.43 •	75.86 ± 1.84 •
	RT-DETR (Zhao et al., 2024)	85.66 ± 2.54 •	83.87 ± 2.67 •	84.73 ± 2.43 •	91.89 ± 2.17 •	80.87 ± 1.88 •	92.04 ± 2.16 •	75.67 ± 1.63 •
	YOLOv9 (Wang et al., 2024b)	83.72 ± 2.04 •	83.14 ± 2.61 •	82.79 ± 1.81 •	91.48 ± 2.27 •	89.46 ± 1.37 •	92.71 ± 1.67 •	76.73 ± 1.87 •
	YOLOv10 (Wang et al., 2024a)	82.63 ± 2.67 •	81.86 ± 2.41 •	82.77 ± 2.47 •	89.97 ± 2.15 •	86.77 ± 2.84 •	92.43 ± 2.17 •	75.72 ± 1.43 •
	TDDet	**92.94 ± 2.53**	**90.76 ± 2.18**	90.21 ± 1.89	**95.68 ± 1.88**	**92.46 ± 2.19**	**95.76 ± 1.06**	**79.63 ± 1.26**
mAP	SSD (Liu et al., 2016)	85.24 ± 3.17 •	83.72 ± 2.87 •	89.72 ± 2.52 •	88.75 ± 2.18 •	87.72 ± 3.26 •	87.92 ± 2.46 •	73.49 ± 1.41 •
	Faster R-CNN (Ren et al., 2016)	86.36 ± 3.06 •	84.57 ± 2.71 •	90.94 ± 2.34 •	90.42 ± 2.40 •	91.28 ± 2.13 •	90.27 ± 2.48 •	74.68 ± 1.57 •
	RetinaNet (Lin et al., 2017b)	85.91 ± 4.95 •	82.95 ± 3.43 •	89.92 ± 3.42 •	91.49 ± 3.16 •	89.39 ± 3.27 •	90.83 ± 2.89 •	73.72 ± 1.82 •
	YOLOv5 (Jocher, 2020)	88.60 ± 3.67 •	86.16 ± 2.75 •	89.46 ± 2.42 •	91.84 ± 2.26 •	91.56 ± 3.13 •	90.38 ± 2.16 •	76.72 ± 1.38 •
	YOLOv8 (Jocher et al., 2023)	90.54 ± 2.73 •	89.14 ± 2.03 •	94.77 ± 2.74 •	94.67 ± 2.77	93.94 ± 2.43 •	**95.47 ± 2.03**	79.87 ± 1.73 •
	DETR (Carion et al., 2020)	89.99 ± 2.68 •	87.87 ± 3.24 •	91.87 ± 2.73 •	93.82 ± 2.13 •	93.42 ± 2.84 •	93.75 ± 2.15 •	77.86 ± 1.54 •
	AX-RetinaNet (Bao et al., 2022)	90.32 ± 2.67 •	88.43 ± 3.13 •	92.79 ± 2.51 •	94.38 ± 2.34 •	92.48 ± 2.94 •	93.95 ± 2.37 •	78.91 ± 1.47 •
	RT-DETR (Zhao et al., 2024)	90.91 ± 3.03 •	87.43 ± 2.72 •	94.72 ± 2.34 •	93.97 ± 2.71 •	94.17 ± 1.88 •	93.75 ± 2.08 •	78.77 ± 1.48 •
	YOLOv9 (Wang et al., 2024b)	91.69 ± 2.63 •	89.14 ± 2.17 •	93.92 ± 1.34 •	94.87 ± 2.17 •	93.67 ± 1.77 •	95.73 ± 1.67	79.39 ± 1.72 •
	YOLOv10 (Wang et al., 2024a)	88.11 ± 2.65 •	86.42 ± 2.45 •	92.97 ± 2.43 •	92.42 ± 2.16 •	91.43 ± 2.46 •	92.48 ± 2.13 •	77.56 ± 1.83 •
	TDDet	**94.33 ± 1.61**	**91.87 ± 2.46**	**96.11 ± 1.46**	**95.54 ± 1.43**	**95.73 ± 2.17**	94.79 ± 1.64	**82.37 ± 1.66**

where TP, FP, and FN are the number of true positives, that of false positives, and that of false negatives, respectively. Precision measures the model's accuracy in correctly identifying tea disease while minimizing FP, ensuring reliable detection and reducing unnecessary interventions. Recall, in contrast, assesses the model's capability to detect all actual disease occurrences, minimizing FN to ensure comprehensive identification of tea diseases.

3.3. Comparative results

To evaluate the performance of TDDet, we conducted a comprehensive comparative analysis against ten main stream object detection models: SSD (Liu et al., 2016), Faster R-CNN (Ren et al., 2016), RetinaNet (Lin et al., 2017b), YOLOv5 (Jocher, 2020), YOLOv8 (Jocher et al., 2023), DETR (Carion et al., 2020), AX-RetinaNet (Bao et al., 2022), RT-DETR (Zhao et al., 2024), YOLOv9 (Wang et al., 2024b), and YOLOv10 (Wang et al., 2024a). Table 2 shows the 5-fold cross-validation results, where "•" indicates that TDDet significantly outperforms other methods, and "∘" denotes that TDDet is significantly outperformed by other methods, based on paired t-tests with a significance level of 0.05. Column "Total" shows the overall results in measure of Precision, Recall and mAP, and others are the corresponding results on each class. The details of each class are discussed in Section 3.1.1.

As shown in the "Total" column of Table 2, our TDDet achieved 91.33% in Precision, 92.94% in Recall, and 94.33% in mAP. In terms of Precision, TDDet outperformed the other models, with improvements of 7.95%, 9.05%, 5.86%, 1.68%, 0.81%, 3.12%, 3.17%, 4.14%, 0.13%, and 9.82% over SSD, Faster R-CNN, RetinaNet, YOLOv5, YOLOv8, DETR, AX-RetinaNet, RT-DETR, YOLOv9, and YOLOv10, respectively. For Recall, TDDet exhibited enhancements of 11.84%, 7.32%, 5.03%, 5.08%, 5.73%, 6.39%, 5.17%, 7.28%, 9.22%, and 10.31%, respectively. Additionally, in mAP, TDDet achieves an improvement of 9.09%, 7.97%, 8.42%, 5.73%, 3.79%, 4.34%, 4.01%, 3.42%, 2.64%, and 6.22% over these models, respectively. These results highlighted TDDet's ability to accurately detect and classify objects across various classes.

From Table 2, TDDet excelled particularly in the TC (95.09%) and TB (91.25%) classes in terms of Precision, where its advanced feature extraction and adaptive mechanisms ensured high accuracy with minimal misclassification. Additionally, TDDet achieved exceptional recall in TRR (95.68%) and TB (92.46%), indicating its ability to detect a large number of objects with a low miss rate, making it highly suitable for applications requiring high sensitivity. The balanced performance of TDDet in TALS (90.87%) and TCLB (88.59%) further highlights its robustness across a variety of detection scenarios, demonstrating strong generalization capabilities.

3.4. Visualization analysis

We visually verified the performance of TDDet in detecting tea diseases. Fig. 11 presents the detection results on randomly selected images using different methods, including TDDet, YOLOv5, YOLOv8, YOLOv9, YOLOv10, and RT-DETR.

From Fig. 11, TDDet demonstrated superior performance in tea disease detection, accurately locating the diseases, and providing clear detection results. In contrast, other models either failed to identify the disease or exhibited suboptimal performance. Specifically, in the first three rows under clear weather and sufficient sunlight, TDDet successfully detected the tea diseases with high confidence scores of 0.80, 0.90, and 0.85, respectively, whereas YOLOv5, YOLO10, and RT-DETR completely failed to detect these diseases. Other methods, such as YOLOv8 and YOLOv9, occasionally detected the diseases with low confidence and precision. In the last two rows, the diffused lighting, likely due to cloud cover, reduced contrast and visibility, causing YOLO10 and RT-DETR to miss the disease in the fourth row. In contrast, TDDet accurately detected the disease with the highest confidence score.

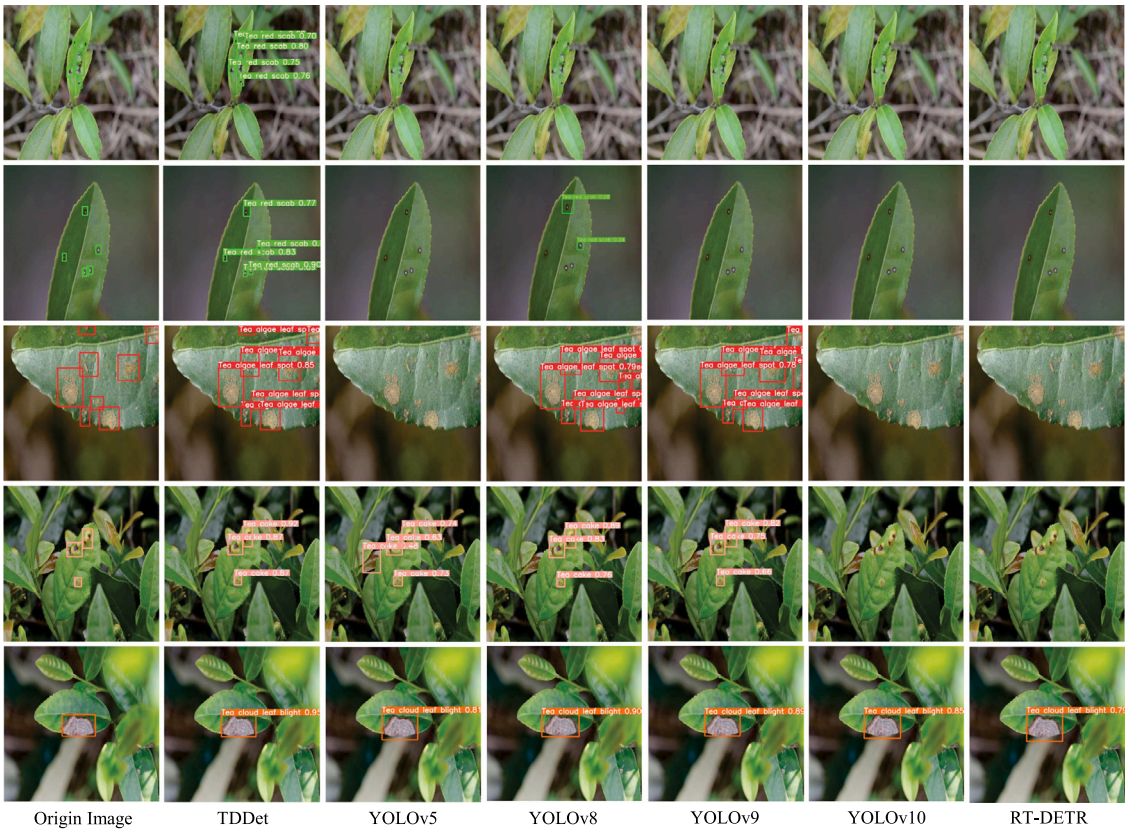

Fig. 11. Visualization of detection results of candidate comparison methods on randomly selected images.

Table 3
Ablation study results for each augmentation method. Hue, BR, RR, NA, and MA represent hue adjustment, brightness adjustment, random rotation, noise addition, and mosaic augmentation, respectively.

Augmentation method					mAP (%)
Hue	BR	RR	NA	MA	
–	–	–	–	–	94.08 ± 1.52
✓	–	–	–	–	94.11 ± 1.29
–	✓	–	–	–	94.17 ± 1.38
–	–	✓	–	–	94.13 ± 1.27
–	–	–	✓	–	94.15 ± 1.11
–	–	–	–	✓	94.19 ± 1.14
✓	✓	✓	✓	✓	**94.33 ± 1.61**

Table 4
Ablation study results for each module.

Module			GFLOPs	mAP (%)	Param (M)
PConv	EMA	DySample			
–	–	–	8.81	92.08 ± 1.21	11.16
✓	–	–	6.65	92.64 ± 1.14	7.55
–	✓	–	9.57	92.87 ± 1.16	12.46
–	–	✓	9.15	92.51 ± 1.63	9.31
✓	✓	–	6.94	93.81 ± 1.57	8.67
✓	–	✓	7.23	93.39 ± 1.42	4.34
✓	✓	✓	7.51	**94.33 ± 1.61**	**5.52**

3.5. Ablation studies

3.5.1. Dataset augmentation ablation

We conducted a series of ablation experiments to evaluate the impact of data augmentations, including hue adjustment, brightness adjustment, random rotation, noise addition, and mosaic augmentation, on our TDDet's performance measured by mAP. More details about these augmentation methods refer to Section 3.1.3. Table 3 summarizes the corresponding results, where "✓" indicates the inclusion of a method, while "–" represents its absence.

Table 3 shows that using hue adjustment alone yielded an mAP of 94.11%, indicating a modest improvement. Brightness adjustment slightly outperformed Hue with an mAP of 94.17%. Random rotation is given an mAP of 94.13%, showing a positive effect but slightly lower than hue and brightness. Noise addition achieved an mAP of 94.15%, providing a small improvement. Mosaic augmentation resulted in the highest individual augmentation mAP of 94.19%, demonstrating its effectiveness. The best performance was achieved when all five augmentation methods were combined, resulting in an mAP of 94.33%.

3.5.2. Architecture ablation

To further analyze the impact of individual components, we conducted a series of ablation experiments by progressively integrating PConv, EMA, and DySample into TDDet. Among them, PConv, and DySample replace the convolution module and upsample module in the basic network. Table 4 summarizes the results, where "✓" indicates the inclusion of a component, while "–" represents its absence.

From Table 4, compared with the baseline network (TDDet without PConv, EMA, and DySample), TDDet with PConv achieved a 0.56% improvement in mAP, while significantly enhancing computational efficiency (in GFLOPs) by 24%. TDDet incorporating both PConv and EMA achieved a 1.17% higher mAP compared to TDDet with PConv. The possible reason is that EMA established interdependencies among spatial positions by utilizing the extensive receptive fields of parallel subnetworks to capture multiscale spatial information. We also observed that TDDet with only DySample outperformed the baseline network, with a 0.43% improvement in mAP. This improvement might be due to the upsampling process being enhanced by adjusting the initial sampling position and offset. In addition, we observed that TDDet with PConv, EMA and DySample achieved the highest mAP (94.33%), demonstrating the effectiveness of these modules.

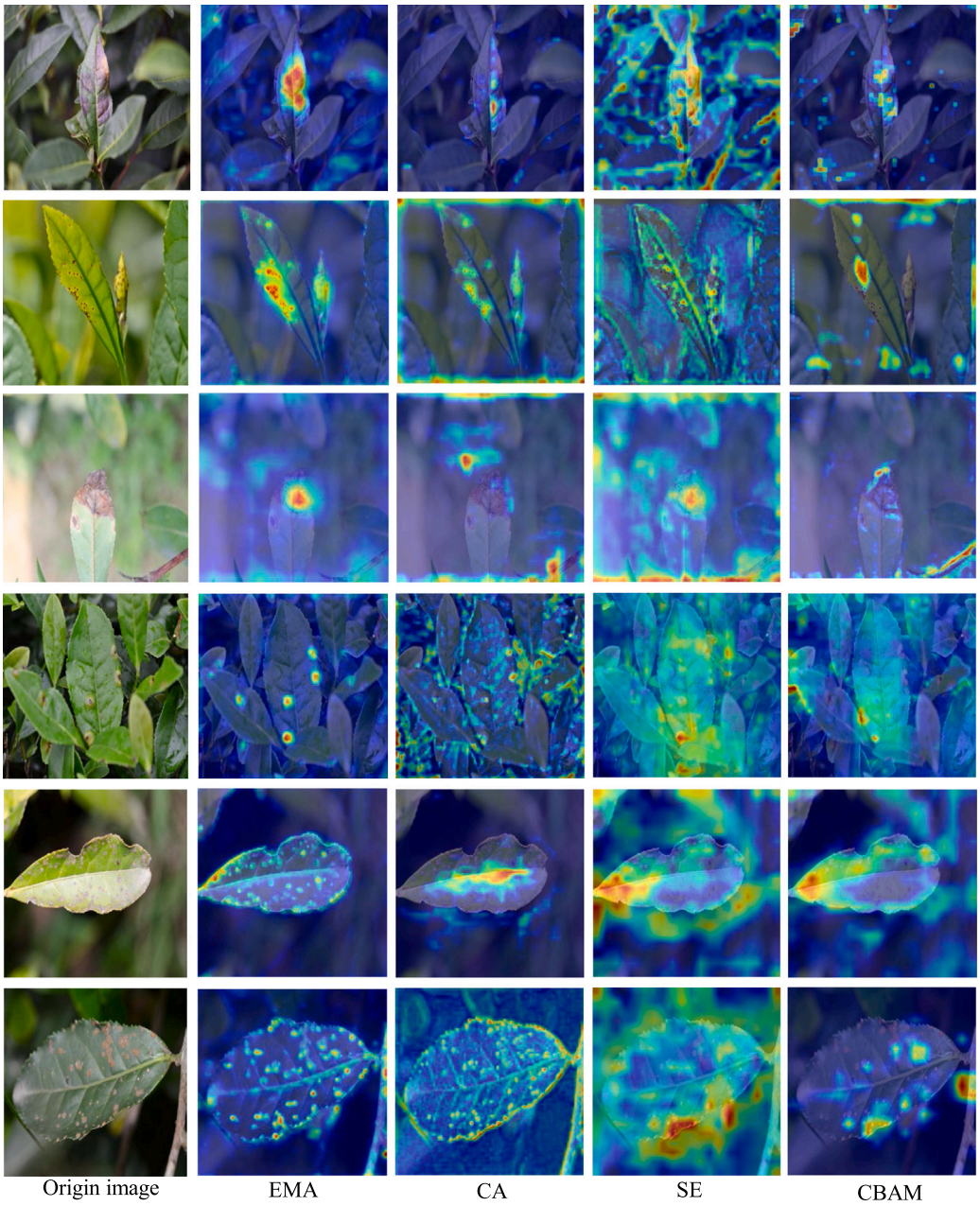

Fig. 12. Comparative heatmaps of the intensity of regions of interest predicted by different attention mechanisms in TDDet.

Table 5
The impact of different attention mechanisms of our TDDet.

Model	P (%)	R (%)	mAP (%)
Baseline	90.37 ± 1.69	91.76 ± 1.84	93.39 ± 1.86
Baseline+CA	90.94 ± 2.07	92.35 ± 1.94	93.83 ± 1.78
Baseline+SE	90.43 ± 1.68	91.97 ± 1.67	93.48 ± 1.73
Baseline+CBAM	90.59 ± 1.83	92.27 ± 1.76	93.75 ± 1.58
Baseline+EMA	**91.33 ± 2.14**	**92.94 ± 2.53**	**94.33 ± 1.61**

3.5.3. Attention mechanism effectiveness

To verify the impact of the proposed EMA, we selected Coordinate Attention (CA) (Hou et al., 2021), Squeeze-and-Excitation (SE) (Hu et al., 2018), and Convolutional Block Attention Module (CBAM) (Woo et al., 2018) as candidate alternatives for comparison. Table 5 shows the corresponding results, where the baseline is our TDDet without EMA (removing EMA from Fig. 3(e)).

From Table 5, the baseline+EMA (our TDDet) outperformed baseline with other attention mechanisms, achieving the highest precision (91.33%), recall (92.94%), and mAP (94.33%). Compared to the baseline, baseline+EMA improved precision by 0.96%, recall by 1.18%, and mAP by 0.94%, demonstrating its effectiveness in enhancing feature expression. In addition, baseline+EMA surpassed the baseline with CA, SE, or CBAM in terms of precision, recall and mAP.

These quantitative results aligned with the visual analysis in Fig. 12. Our TDDet (baseline+EMA) demonstrated superior capability in capturing disease details while maintaining a balance between feature enhancement and noise suppression, thus improving the mAP of the baseline. In contrast, baseline with CA or CBAM struggled with precision, potentially overlooking the subtleties of the affected areas, while baseline+SE tended to overemphasize feature enhancement, leading to excessive focus on noisy regions. These observations indicated that EMA was more suitable as the attention mechanism for TDDet compared to other alternatives.

Table 6
Performance comparison of different backbone networks.

Model	P (%)	R (%)	mAP (%)	Param (M)
CSwin (Dong et al., 2022)	84.43 ± 1.04	82.87 ± 1.98	85.21 ± 1.23	15.15
EfficientViT (Liu et al., 2023b)	90.24 ± 1.31	87.71 ± 1.74	91.38 ± 1.73	17.27
Swin (Liu et al., 2021a)	85.69 ± 2.01	89.73 ± 1.76	88.12 ± 1.65	13.13
MobileNetv4 (Qin et al., 2024)	**91.33 ± 2.14**	**92.94 ± 2.53**	**94.33 ± 1.61**	**5.52**

Table 7
Running efficiency for different methods.

Method	Epoch	Training time per epoch (s)	Test time per slice (s)	Training time (h)	Total number of parameters (M)
SSD (Liu et al., 2016)	250	212	8.14	15.28	24.41
Faster R-CNN (Ren et al., 2016)	230	285	12.98	18.23	79.75
RetinaNet (Lin et al., 2017b)	230	196	6.17	12.91	39.43
YOLOv5 (Jocher, 2020)	210	161	7.35	10.21	21.20
YOLOv8 (Jocher et al., 2023)	200	154	6.31	8.89	25.91
DETR (Carion et al., 2020)	300	585	19.86	50.42	41.12
AX-RetinaNet (Bao et al., 2022)	200	159	5.97	9.17	46.54
RT-DETR (Zhao et al., 2024)	180	182	5.52	9.35	36.00
YOLOv9 (Wang et al., 2024b)	170	154	5.25	7.52	25.31
YOLOv10 (Wang et al., 2024a)	190	172	4.96	9.34	24.42
Ours TDDet	150	128	3.81	5.49	5.52

3.5.4. Backbone network effectiveness

We assessed the effectiveness of various backbone networks for TDDet, including MobileNetv4 (Qin et al., 2024), Swin Transformer (Liu et al., 2021a), CSwin Transformer (Dong et al., 2022), and EfficientViT (Liu et al., 2023b). Table 6 presents their comparative performance in terms of precision, recall, mAP, and model size. From Table 6, TDDet (the model with MobileNetv4) achieved the best performance, with a precision of 91.33%, a recall of 92.94%, and a mAP of 94.33%, while maintaining the smallest parameter size (5.52M). In contrast, the model with CSwin performed the worst, with precision (84.43%), recall (82.87%), and mAP (85.21%). The model with Swin Transformer and EfficientViT demonstrated inferior performance compared to that with MobileNetv4. These results confirmed that MobileNetv4 provided the optimal trade-off between computational efficiency and feature extraction quality.

3.5.5. Running efficiency

To evaluate the computational efficiency of our proposed TDDet, we compared TDDet with state-of-the-art methods in terms of training time, testing time, and model size, as shown in Table 7.

From Table 7, TDDet exhibited a remarkably lightweight model with 5.52M parameters, significantly fewer than SSD (24.41M), Faster R-CNN (79.75M), RetinaNet (39.43M), YOLOv5 (21.20M), YOLOv8 (25.91M), DETR (41.12M), AX-RetinaNet (46.54M), RT-DETR (36.00M), YOLOv9 (25.31M), and YOLOv10 (24.42M). This reduction in parameters led to lower computational costs and faster inference; for example, TDDet achieved the shortest test time per slice (3.81 s), compared to SSD (8.14 s), Faster R-CNN (12.98 s), RetinaNet (6.17 s), YOLOv5 (7.35 s), YOLOv8 (6.31 s), DETR (19.86 s), AX-RetinaNet (5.97 s), RT-DETR (5.52 s), YOLOv9 (5.25 s), and YOLOv10 (4.96 s). Furthermore, TDDet converges within just 150 epochs, completing the training in only 5.49 h—significantly faster than other methods. These results demonstrate that TDDet achieves an excellent balance between efficiency and accuracy, making it a highly competitive choice for real-time applications.

3.6. Generalization experiments

To assess the robustness of our TDDet, we conducted generalization experiments on the Bangladesh Tea Leaf Disease Dataset (BTLD) (Soeb et al., 2023), which consists of five classes of tea leaf diseases: Tea mosquito bug, Brown blight, Black rot, Red spider, and Leaf rust, as shown in Fig. 13. The BTLD dataset was randomly divided into training and test sets in a 4:1 ratio, and the preprocessing methods, including hue adjustment, brightness, random rotation, noise addition, and mosaic augmentation, were used for expanding our the training set. Each image was resized to 640 × 640 pixels to ensure compatibility with our TDDet.

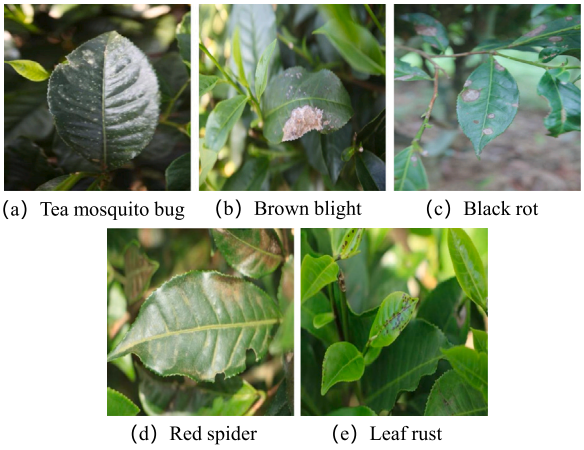

Fig. 13. Disease samples in the BTLD dataset.

Table 8 presents the comparative performance of TDDet with state-of-the-art object detection models. TDDet achieved excellent detection accuracy while maintaining a lightweight architecture (5.52M parameters), with the highest precision, recall, and mAP of 97.46%, 98.08%, and 98.94%, respectively. In addition, TDDet outperformed YOLOv8 in mAP by 1.92%, Faster R-CNN in recall by 3.71%, and RT-DETR in precision by 3.15%, while these models ranked second in their respective metrics.

4. Conclusion

The identification of tea diseases in natural field environments is challenging, requiring timely and accurate detection to mitigate economic losses and ensure effective disease control. This study presents TDDet, a lightweight and efficient tea disease detection model for

Table 8
Performance comparison of different models on BTLD.

Model	P (%)	R (%)	mAP (%)	Param (M)
SSD (Liu et al., 2016)	93.25	94.55	93.65	24.41
Faster R-CNN (Ren et al., 2016)	93.54	94.47	94.74	79.75
RetinaNet (Lin et al., 2017b)	95.35	94.93	93.97	36.43
YOLOv5 (Jocher, 2020)	95.74	89.77	95.72	21.20
YOLOv8 (Jocher et al., 2023)	96.15	96.89	97.02	25.91
DETR (Carion et al., 2020)	94.73	93.58	95.91	41.12
AX-RetinaNet (Bao et al., 2022)	94.82	94.73	94.92	46.54
RT-DETR (Zhao et al., 2024)	94.31	93.45	95.27	36.00
YOLOv9 (Wang et al., 2024b)	96.13	90.76	96.46	25.31
YOLOv10 (Wang et al., 2024a)	92.45	93.76	94.96	24.42
Our TDDet	**97.46**	**98.18**	**98.94**	**5.52**

real-world agricultural applications. TDDet leverages Depthwise Separable Convolutions (DW) and the Multi-Query Attention (MQA) mechanism to capture intricate disease patterns while maintaining computational efficiency. A novel Cross-scale Feature Fusion (CFF) module improves feature aggregation by merging multiscale information through bidirectional pathways, enhancing feature complementarity and boosting the model's ability to capture detailed disease characteristics. Meanwhile, the Dysample module is employed to dynamically adjust the upsampling rate to balance detection accuracy and computational cost. Experimental results demonstrate that TDDet achieves state-of-the-art performance with significantly fewer parameters, enabling real-time deployment on mobile platforms for precise and timely disease management.

Many natural factors, such as climate (temperature, humidity, lighting) and the complex background (vegetation, soil, leaf occlusion), influence the performance of tea disease detection models. However, the experimental dataset was collected only under natural daylight conditions at fixed points, limiting diversity and failing to cover extreme environments (e.g., rain, dense occlusions), potentially leading to an incomplete robustness evaluation. In addition, as shown in Table 2, TDDet achieves the highest recall (95.76%) in the TRR class but with a limited margin over YOLOv9 and RT-DETR, while its precision is not the highest, indicating a trade-off between recall and false positives.

One future effort will focus on collecting datasets that encompass diverse environmental conditions, enabling the validation of the TDDet model's performance and enhancing its robustness across various applications. Another work will incorporate adaptive mechanisms, including refining decision boundaries and integrating adaptive thresholding, to further improve TDDet's precision and recall in the TRR class while enhancing its generalization and robustness across diverse scenarios.

CRediT authorship contribution statement

Yange Sun: Writing – review & editing, Investigation, Formal analysis, Conceptualization. **Zhihao Li:** Writing – original draft, Visualization, Software, Data curation. **Huaping Guo:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Yan Feng:** Writing – review & editing, Supervision. **Yongqiang Tang:** Supervision, Resources. **Wensheng Zhang:** Supervision, Formal analysis. **Jingqiu Gu:** Supervision.

Funding

The study was funded in part by Henan Province Key Research and Development Project under Grant 252102220046, in part by Henan Joint Fund for Science and Technology Research under Grant 20240012, in part by Key Scientific Research Projects of Higher Education Institutions in Henan Province under Grant 25B520004, and in part by the Open Fund of the Engineering Research Center of Intelligent Swarm Systems, Ministry of Education under Grant ZZU-CISS-2024004.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our code and dataset are available at https://github.com/hpguo1982/TDDet.

References

Ashok, S., Kishore, G., Rajesh, V., Suchitra, S., Sophia, S.G., Pavithra, B., 2020. Tomato leaf disease detection using deep learning techniques. In: 2020 5th International Conference on Communication and Electronics Systems. ICCES, IEEE, pp. 979–983.
Bag, S., Mondal, A., Banik, A., 2022. Exploring tea (Camellia sinensis) microbiome: Insights into the functional characteristics and their impact on tea growth promotion. Microbiol. Res. 254, 126890.
Bala, R., Sharma, A., Goel, N., 2024. Comparative analysis of diabetic retinopathy classification approaches using machine learning and deep learning techniques. Arch. Comput. Methods Eng. 31 (2), 919–955.
Bao, W., Fan, T., Hu, G., Liang, D., Li, H., 2022. Detection and identification of tea leaf diseases based on AX-RetinaNet. Sci. Rep. 12 (1), 2183.
Bauters, J.F., Heck, M.J., John, D., Dai, D., Tien, M.-C., Barton, J.S., Leinse, A., Heideman, R.G., Blumenthal, D.J., Bowers, J.E., 2011. Ultra-low-loss high-aspect-ratio Si3N4 waveguides. Opt. Express 19 (4), 3163–3174.
Bhuyan, P., Singh, P.K., Das, S.K., 2024. Res4net-CBAM: a deep cnn with convolution block attention module for tea leaf disease diagnosis. Multimedia Tools Appl. 83 (16), 48925–48947.
Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2020. Global context networks. IEEE Trans. Pattern Anal. Mach. Intell. 45 (6), 6881–6895.
Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.
Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B., 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134.
Gallian, J.A., 2012. Graph labeling. Electron. J. Comb. DS6–Dec.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377.
Gyapong, J.O., Remme, J.H., 2001. The use of grid sampling methodology for rapid assessment of the distribution of bancroftian filariasis. Trans. R. Soc. Trop. Med. Hyg. 95 (6), 681–686.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
He, J., Zhang, S., Yang, C., Wang, H., Gao, J., Huang, W., Wang, Q., Wang, X., Yuan, W., Wu, Y., et al., 2024. Pest recognition in microstates state: an improvement of YOLOv7 based on spatial and channel reconstruction convolution for feature redundancy and vision transformer with Bi-Level Routing Attention. Front. Plant Sci. 15, 1327237.
Hossain, S., Mou, R.M., Hasan, M.M., Chakraborty, S., Razzak, M.A., 2018. Recognition and detection of tea leaf's diseases using support vector machine. In: 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications. CSPA, IEEE, pp. 150–154.
Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Hu, G., Wang, H., Zhang, Y., Wan, M., 2021. Detection and severity analysis of tea leaf blight based on deep learning. Comput. Electr. Eng. 90, 107023.

Jocher, G., 2020. YOLOv5 by Ultralytics. http://dx.doi.org/10.5281/zenodo.3908559, URL https://github.com/ultralytics/yolov5.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO. URL https://github.com/ultralytics/ultralytics.

Krishnakumar, V., Kumar, T.R., Murugesan, P., 2024. Tea (Camellia sinensis (L.) O. Kuntze). In: Soil Health Management for Plantation Crops: Recent Advances and New Paradigms. Springer, pp. 391–486.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.

Li, H., Shi, H., Du, A., Mao, Y., Fan, K., Wang, Y., Shen, Y., Wang, S., Xu, X., Tian, L., et al., 2022. Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-RNet. Front. Plant Sci. 13, 922797.

Li, L., Zhao, Y., 2025. Tea disease identification based on ECA attention mechanism ResNet50 network. Front. Plant Sci. 16, 1489655.

Liang, J., Liang, R., Wang, D., 2025. A novel lightweight model for tea disease classification based on feature reuse and channel focus attention mechanism. Eng. Sci. Technol. Int. J. 61, 101940.

Lin, J., Bai, D., Xu, R., Lin, H., 2023. TSBA-YOLO: An improved tea diseases detection model based on attention mechanisms and feature fusion. Forests 14 (3), 619.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37.

Liu, S., Chen, J., He, S., Shi, Z., Zhou, Z., 2023a. Few-shot learning under domain shift: Attentional contrastive calibrated transformer of time series for fault diagnosis under sharp speed variation. Mech. Syst. Signal Process. 189, 110071.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020. Deep learning for generic object detection: A survey. Int. J. Comput. Vis. 128, 261–318.

Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y., 2023b. Efficientvit: Memory efficient vision transformer with cascaded group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14430.

Liu, Y., Pu, H., Sun, D.-W., 2021b. Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. Trends Food Sci. Technol. 113, 193–204.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al., 2024. MobileNetV4-Universal models for the mobile ecosystem. arXiv preprint arXiv:2404.10518.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 658–666.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883.

Soeb, M.J.A., Jubayer, M.F., Tarin, T.A., Al Mamun, M.R., Ruhad, F.M., Parven, A., Mubarak, N.M., Karri, S.L., Meftaul, I.M., 2023. Tea leaf disease detection and identification based on YOLOv7 (YOLO-t). Sci. Rep. 13 (1), 6078.

Sun, C., Huang, C., Zhang, H., Chen, B., An, F., Wang, L., Yun, T., 2022. Individual tree crown segmentation and crown width extraction from a heightmap derived from aerial laser scanning data using a deep learning framework. Front. Plant Sci. 13, 914974.

Sun, Y., Jiang, Z., Zhang, L., Dong, W., Rao, Y., 2019. SLIC_SVM based leaf diseases saliency map extraction of tea plant. Comput. Electron. Agric. 157, 102–109.

Sun, Y., Wu, F., Guo, H., Li, R., Yao, J., Shen, J., 2023. TeaDiseaseNet: multi-scale self-attentive tea disease detection. Front. Plant Sci. 14, 1257212.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2023a. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7464–7475.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G., 2024a. Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D., 2019. Carafe: Content-aware reassembly of features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3007–3016.

Wang, Y., Xu, R., Bai, D., Lin, H., 2023b. Integrated learning-based pest and disease detection method for tea leaves. Forests 14 (5), 1012.

Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M., 2024b. Yolov9: Learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616.

Wei, Y., Wen, Y., Huang, X., Ma, P., Wang, L., Pan, Y., Lv, Y., Wang, H., Zhang, L., Wang, K., et al., 2024. The dawn of intelligent technologies in tea industry. Trends Food Sci. Technol. 104337.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Xia, E.-H., Tong, W., Wu, Q., Wei, S., Zhao, J., Zhang, Z.-Z., Wei, C.-L., Wan, X.-C., 2020. Tea plant genomics: achievements, challenges and perspectives. Hortic. Res. 7.

Xia, Y., Yuan, W., Zhang, S., Wang, Q., Liu, X., Wang, H., Wu, Y., Yang, C., Xu, J., Li, L., et al., 2024. Classification and identification of tea diseases based on improved YOLOv7 model of MobileNeXt. Sci. Rep. 14 (1), 11799.

Xiong, H., Li, J., Wang, T., Zhang, F., Wang, Z., 2024. EResNet-SVM: an overfitting-relieved deep learning model for recognition of plant diseases and pests. J. Sci. Food Agric. 104 (10), 6018–6034.

Xu, Q., Yang, Y., Hu, K., Chen, J., Djomo, S.N., Yang, X., Knudsen, M.T., 2021. Economic, environmental, and emergy analysis of China's green tea production. Sustain. Prod. Consum. 28, 269–280.

Xue, X., Jin, S., An, F., Zhang, H., Fan, J., Eichhorn, M.P., Jin, C., Chen, B., Jiang, L., Yun, T., 2022. Shortwave radiation calculation for forest plots using airborne LiDAR data and computer graphics. Plant Phenom..

Xue, Z., Xu, R., Bai, D., Lin, H., 2023. YOLO-tea: A tea disease detection model improved by YOLOv5. Forests 14 (2), 415.

Yang, Z., Feng, H., Ruan, Y., Weng, X., 2023. Tea tree pest detection algorithm based on improved Yolov7-Tiny. Agriculture 13 (5), 1031.

Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q., 2021. Rethinking rotated object detection with gaussian wasserstein distance loss. In: International Conference on Machine Learning. PMLR, pp. 11830–11841.

Ye, R., Gao, Q., Li, T., 2024a. BRA-YOLOv7: improvements on large leaf disease object detection using FasterNet and dual-level routing attention in YOLOv7. Front. Plant Sci. 15, 1373104.

Ye, R., Shao, G., He, Y., Gao, Q., Li, T., 2024b. YOLOv8-RMDA: Lightweight YOLOv8 network for early detection of small target diseases in tea. Sensors 24 (9), 2896.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T., 2016. Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 516–520.

Zhang, Y., Li, X., Wang, M., Xu, T., Huang, K., Sun, Y., Yuan, Q., Lei, X., Qi, Y., Lv, X., 2024. Early detection and lesion visualization of pear leaf anthracnose based on multi-source feature fusion of hyperspectral imaging. Front. Plant Sci. 15, 1461855.

Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T., 2022. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 506, 146–157.

Zhang, Q.-L., Yang, Y.-B., 2021. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2235–2239.

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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16965–16974.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 12993–13000.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 13001–13008.

Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W., 2023. Biformer: Vision transformer with bi-level routing attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10323–10333.