Detection of Wheat Pest and Disease in Complex Backgrounds Based on Improved YOLOv8 Model

Dandan Zhong[1], Penglin Wang*[2], Jie Shen*[3], Dongxu Zhang*[4]
College of Agriculture, Shanxi Agricultural University, Jinzhong, China[1]
School of Electronics and Information Engineering, Anhui Jianzhu University, Hefei, China[2]
Changzhi University, Changzhi, China[3]
Millet Research Institute, Shanxi Agricultural University, Key Laboratory of Sustainable Dryland Agriculture,
(Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Changzhi, China[4]

***Abstract*—Detecting wheat diseases and pests, particularly those characterized by small targets amidst complex background interference, presents a significant challenge in agricultural research. To address this issue and achieve precise and efficient detection, we propose an enhanced version of YOLOv8, termed MGT-YOLO, which incorporates multi-scale edge enhancement and visual remote dependency mechanisms. Our methodology begins with the creation of a comprehensive dataset, WheatData, comprising 2393 high-resolution images capturing various wheat diseases and pests across different growth stages in diverse agricultural settings. To improve the detection of small targets, we implemented a multi-scale edge amplification technique within the backbone network of YOLOv8, enhancing its ability to capture minute details of wheat diseases and pests. Furthermore, we introduced the C2f_GlobalContext module in the neck network, which integrates global contextual relationships and facilitates the fusion of features from small-sized objects by leveraging remote dependencies in visual imagery. Additionally, we incorporated a Vision Transformer module into the neck network to enhance the processing efficiency of small-scale disease and pest features. The proposed MGT-YOLO network was rigorously evaluated on the WheatData dataset. The results demonstrated significant improvements, with mAP@0.5 values of 90.0% for powdery mildew and 65.5% for smut disease, surpassing the baseline YOLOv8 by 5.3% and 6.8%, respectively. The overall mAP@0.5 reached 89.5%, representing a 2.0% improvement over YOLOv8 and outperforming other state-of-the-art detection methods. These findings suggest that MGT-YOLO is a promising solution for real-time detection of agricultural diseases and pests, offering enhanced accuracy and efficiency in complex agricultural environments.**

***Keywords*—Wheat disease and pest; YOLOv8; edge amplification; visual remote dependency; global context; vision transformer**

I. Introduction

Detecting pests and diseases in wheat is a vital component of agricultural production, playing a vital role in ensuring wheat quality. During the cultivation of wheat, factors such as climate and geographical conditions can lead to varying levels of interference from diseases and pests [1]. These issues not only compromise wheat quality but also disrupt normal agricultural operations [2], [3]. Accurate and timely identification of these diseases and pests during cultivation can mitigate potential problems to a significant extent [4], [5].

In the field of computer vision-based object detection, two primary architectural paradigms have emerged: two-stage and single-stage detection models. Representative two-stage detectors including RCNN [6], Fast-RCNN [7], and Faster-RCNN [8] are characterized by their hierarchical processing architecture that achieves superior localization precision and detection accuracy. However, these models suffer from inherent computational complexity due to their region proposal generation mechanism, resulting in suboptimal inference speeds that limit their practical applicability in real-time agricultural disease and pest monitoring scenarios. By contrast, single-stage detection frameworks represented by SSD [9] and the YOLO series [10], [11] employ end-to-end detection pipelines that directly predict bounding boxes and class probabilities. This architectural simplification enables these models to achieve a favorable trade-off between detection performance and computational efficiency, making them particularly suitable for real-time agricultural applications. Despite significant advancements in detection accuracy through successive iterations, current YOLO variants still exhibit limitations in recognizing small-scale pathogenic features under complex field conditions with cluttered backgrounds [12], [13], an inherent challenge exacerbated by the scale variations and occlusion patterns typical in agricultural environments.

To tackle the inherent challenges of YOLO architectures in capturing and integrating fine-grained features across backbone and neck network hierarchies, this research presents an innovative Multi-scale Edge Augmentation Framework (MEAM) specifically tailored for improved detection of minute wheat disease patterns and pest characteristics. This architecture-level enhancement strategically reinforces feature representation through multi-level edge preservation operations. Additionally, a feature fusion module named C2f_GlobalContext is introduced to capture global contextual relationships and strengthen the fusion of small-object features by leveraging long-range dependencies in visual images. Furthermore, the efficiency advantages of the Vision Transformer network are utilized to improve the processing of small-scale disease and pest features.

Thus, this study presents the MGT-YOLO network, which aims to achieve precise and rapid detection of wheat diseases and pests, addressing the challenges of small-object detection in agricultural applications.

As summarized above, the key contributions can be described as follows:

*Corresponding authors

1. Proposed the MGT-YOLO approach for the detection of small-scale wheat diseases and pests. This method achieves higher precision and lightweight performance compared to traditional models.

2. Designed and integrated the Multi-scale Edge Augmentation Mechanism (MEAM) into the backbone network to enhance the extraction of fine-grained features, such as wheat disease and pest characteristics, from images.

3. Developed the C2f_GlobalContext feature fusion module, which incorporates global contextual relationships to strengthen the fusion of features for small-scale diseases and pests in images. This module enhances feature integration by capturing long-range dependencies in visual images. Additionally, the Vision Transformer module was introduced into the neck network to improve the efficiency of processing small-scale disease and pest features.

The structure of this paper is organized as follows: First, we introduce the related work of computer vision detection technology in agricultural pest and disease detection. Then, we present the improvements made based on the YOLOv8 algorithm in feature extraction and feature fusion, as well as the overall workflow of the proposed algorithm framework, MGT-YOLO. Next, we describe the experimental work on wheat pest and disease detection, including the self-constructed dataset WheatData, the evaluation metrics used in the experiments, a comparison of the proposed MGT-YOLO algorithm with other state-of-the-art algorithms, and the results of ablation studies. Finally, we summarize the experimental findings and provide an outlook for future research.

II. Related Work

The application prospects of computer vision technology in the agricultural field are vast [14], [15]. Quan [16] and colleagues employed an improved Faster R-CNN model to detect maize diseases in complex field environments. As a two-stage detection framework, Faster R-CNN exhibits inherent computational latency that fails to satisfy the stringent real-time processing demands characteristic of modern agricultural robotics and automated crop monitoring systems. This limitation primarily stems from its region proposal network architecture and sequential feature processing pipeline, which significantly constrain inference speed in field deployment scenarios. Liangquan [17] and Jizhong Deng [18] used an improved YOLOv7 model to detect rice pests and diseases by replacing the YOLO backbone with lightweight networks such as MobileNetV3 or GhostNet. While this approach improved real-time detection performance, it did not effectively enhance detection accuracy when the base model already satisfied real-time requirements. Similarly, Yinkai [19] implemented a self-attention mechanism in the YOLOv8 backbone to detect tea pests and diseases. Although this method improved feature extraction capabilities, it introduced a significant number of parameters and required extensive exploration to determine the optimal placement of the attention mechanism.

Wang [20] integrated the Global Attention Mechanism (GAM) into the C2f structure of YOLOv8's backbone network, enabling the model to better comprehend the overall semantics of the image. Zhang [21] designed the C2f_ODConv module, introducing it alongside ODConv into YOLOv8's backbone network, enhancing feature extraction capabilities while reducing parameter redundancy through a multi-dimensional attention mechanism. Qu [22] replaced the convolutional modules in YOLOv8's backbone network with spatial depth convolutions. Zhen [23] further strengthened YOLOv8's feature representation capabilities by introducing the Multi-Scale Feature Attention Module (MSFAM). Luo [24] enhanced YOLOv8's ability to capture fine details and its detection accuracy by incorporating Channel-Priority Attention Dynamic Snake Convolution and a Dynamic Small Object Detection Head Layer (DyHead-SODL). Although these efforts have enhanced the feature extraction capability of the backbone network to some extent, they have significantly increased the number of parameters in the backbone network. Wang [25] enhanced the feature extraction capability of the base model by incorporating their self-designed PotentNet network into the backbone of YOLOv8. However, this strategy did not account for long-range dependencies between different features, indicating that there remains significant potential for improving the base model's feature extraction ability.

Zhengyu Zhang [26] and colleagues incorporated Coordinate Attention (CA) and lightweight GSConv into YOLOv8 to minimize the model's parameters and enhance feature extraction in the backbone to some extent. However, during the prediction stage, the performance relied heavily on the feature fusion capability of the neck network. As a result, the method was insufficient for detecting small-scale agricultural pests and diseases. Bai Shao [27] and colleagues enhanced the feature fusion capabilities of YOLOv8 by introducing a multi-head attention mechanism for tea pest and disease detection. While this approach improved feature integration, it significantly increased computational demands and model parameter count [28], making it less suitable for resource-constrained inference devices. Therefore, improving feature fusion capabilities while maintaining model lightweightness remains a critical consideration [29].

From the above works, it is evident that convolutional neural network-based teams often focus on enhancing the lightweight design of backbones and improving feature extraction for crop pest and disease detection tasks. However, relatively little attention has been given to optimizing the fusion of extracted features. Additionally, the lightweight design of feature fusion networks has not been sufficiently addressed.

To systematically address these challenges, this study introduces a comprehensive algorithmic refinement framework for YOLO-series architectures, focusing on optimizing the model's capability in multi-scale feature extraction and hierarchical fusion mechanisms specifically for small-sized agricultural pest and disease patterns. The proposed improvements span both backbone feature representation learning and neck network feature integration modules, while maintaining computational efficiency through lightweight structural optimizations.

In terms of base model selection, YOLOv8 [30] is an algorithm in the field of object detection that excels in both lightweight design and detection performance. However, based on the analysis of related improvements to YOLOv8, it is evident that YOLOv8 still has several shortcomings, such as room for enhancement in feature extraction and feature fusion. Therefore, this paper chooses YOLOv8 as the base model and explores improvements in feature extraction and feature fusion.

III. METHODS

A. Multi-Scale Edge Amplification Module

The backbone of the YOLOv8 performs layer-by-layer feature extraction through multiple convolutional layers. However, when dealing with multi-scale small-object features, it still suffers from insufficient feature extraction capabilities [31]. Inspired by the initial block design of DEM [32], we made modifications to adapt it for real-time detection tasks, enhancing the ability to capture features across multiple scales. This enhanced module is referred to as the Multi-scale Edge Augmentation Mechanism (MEAM).

The structure of MEAM, shown in Fig. 1, comprises an AP (Average Pooling) layer with a 3*3 kernel, a Conv (Convolution) layer with a 1*1 kernel, and an EE (Edge Enhancer) module. The EE module itself is composed of an AP layer with a 3*3 kernel and a Conv layer with a 1×1 kernel. By leveraging residual connections, the EE module performs deep extraction of input features to capture object edges in the feature maps.

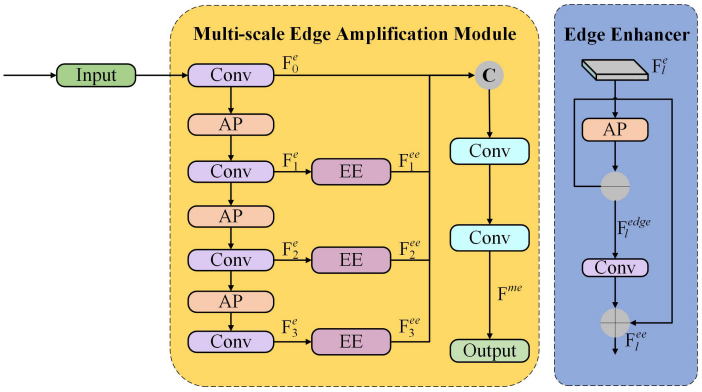

Fig. 1. Schematic diagram of the MEAM module structure.

When input features are processed through the MEAM module, the following steps are performed: First, the input is subjected to a 1*1 convolution operation to produce the feature map $\mathbf{F}_0^e$. Subsequently, $\mathbf{F}_1^e$, $\mathbf{F}_2^e$, and $\mathbf{F}_3^e$ are obtained through three successive average pooling and convolution operations. These features are then passed through the EE module to yield enhanced features $\mathbf{F}_1^{ee}$, $\mathbf{F}_2^{ee}$, and $\mathbf{F}_3^{ee}$. The enhanced features, together with Fe0, are concatenated along the channel dimension. Finally, two additional convolution operations are applied to the concatenated features to produce Fme, which is passed to the neck network for subsequent computations.

The mathematical operations involved in processing feature information through MEAM are described in Eq. (1) to (6). In these equations, ϕ_{1*1} represents convolution operations using a Conv layer with a 1*1 kernel, and AP represents average pooling operations with a 3*3 kernel. $\mathbf{F}^{local}$ represents the input feature.

$$\mathbf{F}_0^e = \phi_{1*1}(\mathbf{F}^{local}) \tag{1}$$

$$\mathbf{F}_{t+1}^e = AP(\phi'_{1\times1}(\mathbf{F}_t^e)), (0 \le t \le 2) \tag{2}$$

$$\mathbf{F}_l^{ee} = \psi(\mathbf{F}_l^e), (1 \le l \le 3) \tag{3}$$

$$\mathbf{F}_l^{\text{edge}} = \mathbf{F}_l^e - AP(\mathbf{F}_l^e) \tag{4}$$

$$\mathbf{F}_l^{ee} = \phi'_{1\times1}(\mathbf{F}_l^{\text{edge}}) + \mathbf{F}_l^e \tag{5}$$

$$\mathbf{F}^{me} = \phi_{1\times1}([\mathbf{F}_0^e, \mathbf{F}_1^{ee}, \mathbf{F}_2^{ee}, \mathbf{F}_3^{ee}]) \tag{6}$$

B. C2f_GlobalContext for Capturing Visual Remote Dependencies

The neck architecture in YOLOv8 demonstrates suboptimal performance in handling multi-scale feature flows, particularly for capturing discriminative patterns of small-object disease manifestations and pest morphological characteristics. This limitation leads to compromised feature fusion efficacy in cross-scale aggregation. To address this critical bottleneck, we propose the integration of a Global Context (GC) mechanism, an attention-based architectural enhancement that establishes long-range dependency modeling across hierarchical feature representations [33]. This strategic modification enables contextual reasoning over global receptive fields while preserving local structural details essential for fine-grained pest and disease recognition. A new module, C2f_GlobalContext, incorporating the GC mechanism, was designed to replace specific layers of the network's original C2f module.

The structure of the GC mechanism, shown in Fig. 2, consists of a convolutional layer (Conv) with a 1*1 kernel, a Softmax layer, and a Layer Normalization (LayerNorm) layer. The processing flow of the GC mechanism is described in Eq. (8). When input features x are passed into the GC mechanism, they first undergo W_k processing in the ContextModeling module, where features are aggregated using a weighted average with weights α_j (calculated as shown in Eq. (7)). This step groups the features from all positions to generate global context features v_1. The v_1 features are then processed through the Transform layer, which includes W_{v_1} convolution, LayerNorm, and W_{v_2} convolution in sequence. These operations capture channel dependencies to produce refined global context features v_2. Finally, the global context features v_2 are aggregated with the input features x.

$$\alpha_j = \frac{e^{W_k \mathbf{x}_j}}{\sum_m e^{W_k \mathbf{x}_m}} \tag{7}$$

$$\mathbf{z} = \mathbf{x} + W_{v2}\text{ReLU}\left(\text{LN}\left(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k \mathbf{x}_j}}{\sum_{m=1}^{N_p} e^{W_k \mathbf{x}_m}}\mathbf{x}_j\right)\right) \tag{8}$$

By capturing long-range visual dependencies, this approach enriches the gradient flow of small-object features, significantly enhancing the feature fusion capability of the neck network.

As illustrated in Fig. 3, the C2f_GlobalContext module is composed primarily of CBS units and GCBottleneck units.

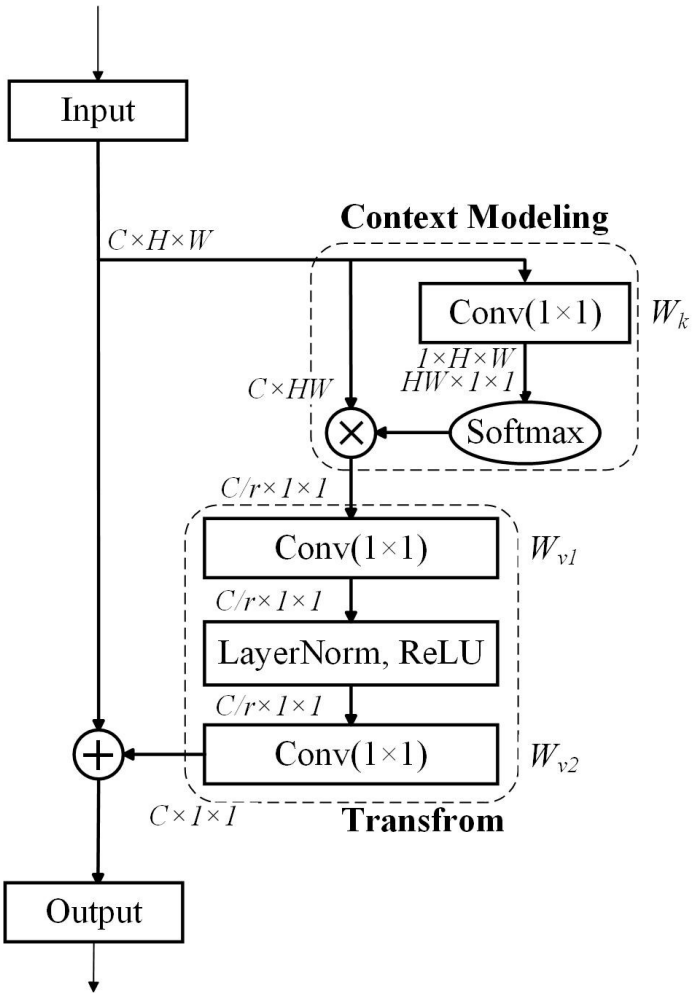

Fig. 2. Schematic diagram of the GlobalContext network structure.

Each GCBottleneck unit integrates a CBS unit with a GlobalContext unit, enabling the module to extract features from the input data across multiple hierarchical levels and varying degrees of abstraction. These features are subsequently fused through element-wise addition, resulting in a comprehensive and robust integration.

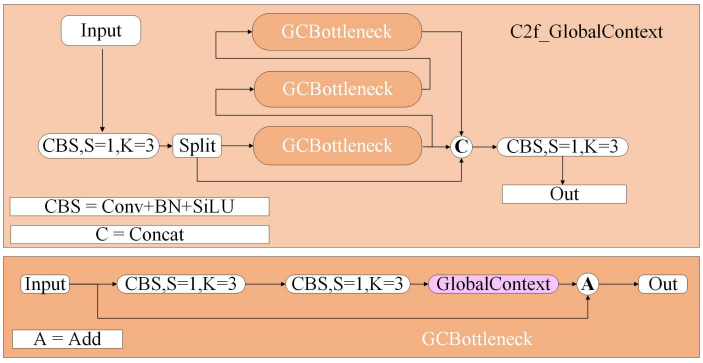

Fig. 3. Structure diagram of C2f_GlobalContext module.

To demonstrate the enhanced feature fusion capacity of the C2f_GlobalContext module, we conducted a controlled comparison of activation patterns between the baseline C2f module and our proposed architecture using the Wheat-Data dataset. Fig. 4 systematically presents this analysis: panel (a) displays object detection outputs from both architectures, while panels (b) and (c) contrast intermediate feature representations extracted from equivalent network depths in the C2f and C2f_GlobalContext models respectively.

Comparative analysis of Fig. 4 reveals that the network incorporating the C2f_GlobalContext module achieves marked improvement in feature recognition accuracy. Specifically, this enhanced architecture exhibits enhanced precision in localizing wheat powdery mildew-related features while effectively suppresses extraneous background interference. Conversely, the baseline C2f module not only fails to accurately delineate disease-specific characteristics but also demonstrates pronounced susceptibility to background artifacts, as evidenced by its inappropriate attention allocation to non-pathological regions.

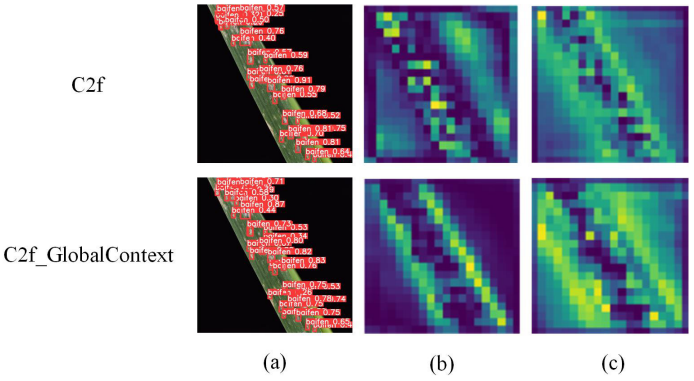

Fig. 4. Feature visualization comparison of C2f and C2f_GlobalContext modules on the WheatData.

C. Architectural Design of the MGT-YOLO

The architectural configuration of the MGT-YOLO network is visually presented in Fig. 5. Our methodology extends the YOLOv8 framework through three strategic enhancements: Implementation of a lightweight Multi-scale Enhancement Attention Module (MEAM) at the backbone's terminal layer, specifically engineered to amplify discriminative feature representation through cross-channel interactions; Substitution of the standard C2f module with a Global Context-aware C2f variant (C2f_GlobalContext) in the neck network, enhancing multi-scale feature fusion through spatial-channel contextual modeling; Integration of a Vision Transformer (ViT) layer [34] with adaptive window attention, strategically positioned in the neck architecture to address the critical challenge of capturing long-range dependencies among fragmented pest and disease patterns, particularly beneficial for small-object feature preservation.

The operational pipeline of MGT-YOLO for detecting wheat pest and disease features in digital images comprises two principal phases. During the preprocessing stage, input images undergo dimension standardization through bilinear interpolation to achieve a fixed resolution of 640*640 pixels. Subsequently, the architecture's backbone network employs

a hierarchical feature extraction mechanism, utilizing convolutional blocks to progressively capture multi-scale feature representations - from low-level texture patterns to high-level semantic information - through depthwise separable convolution operations. Following the feature extraction phase, the backbone network sequentially delivers multi-level feature representations (low, medium, and high-resolution) to the neck network for hierarchical feature fusion. Through bidirectional cross-scale connections, the neck network systematically propagates these enhanced feature maps across three distinct detection scales to the head network. Ultimately, the detection head generates precise bounding box coordinates and category probability distributions by simultaneously analyzing the complementary spatial and semantic information contained in the multi-scale feature maps.

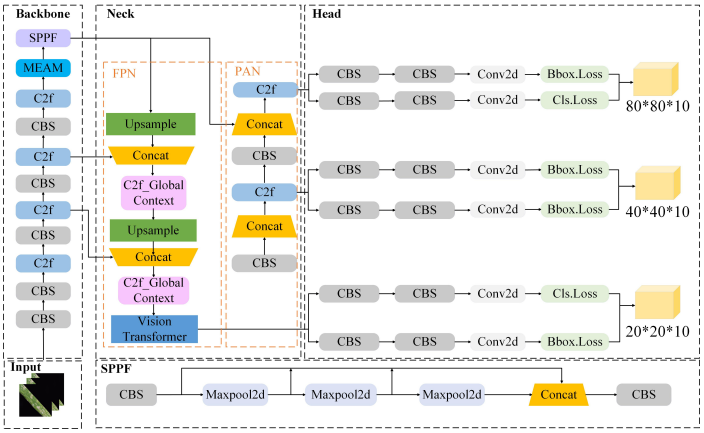

Fig. 5. Structure of MGT-YOLO network.

The head network performs dual-task optimization by simultaneously computing classification and localization losses, which are subsequently minimized through the Stochastic Gradient Descent (SGD) optimizer. The classification branch employs Binary Cross-Entropy (BCE) to quantify prediction errors, while the localization module adopts the Complete Intersection over Union ($CIoU$) Loss [35] for bounding box regression, as formalized in Eq. (9). Here, b and b^{gt} denote the geometric center coordinates of the predicted and ground-truth bounding boxes, respectively. $p^2(b, b^{gt})$ computes the Euclidean distance between the two centers, and the Intersection over Union (IoU) measures the intersection-over-union ratio between the predicted and ground-truth boxes. The model incorporates two critical parameters: the weight coefficient α and the consistency coefficient v. The IoU metric is mathematically formulated in Eq. (10), where A and B denote the predicted bounding box and ground-truth box, respectively. This metric quantifies spatial overlap by calculating the ratio between the area of intersection and the area of union of the two boxes. The derivation of coefficients α and v follows distinct computational procedures as specified in Eq. (11) and (12), respectively. In Eq. (12), w,h and w^{gt},h^{gt} represent the width and height parameters of the predicted and ground-truth boxes, respectively.

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{9}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

$$\alpha = \frac{v}{(1 - loU) + v} \tag{11}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{12}$$

IV. Experiment

The study utilizes MGT-YOLO for training, validation, and testing on the Wheat-Data dataset. Additionally, the detection performance of MGT-YOLO is compared with models from related studies, followed by a comprehensive data analysis.

A. Dataset Specifications and Experimental Configuration

The study focuses on a custom-built wheat pest and disease detection dataset, Wheat-Data, which comprises 2393 images of six types of wheat pests and diseases. An 8:1:1 split ratio is implemented for the dataset allocation across training, validation, and test subsets respectively. The images were manually captured at different stages of wheat growth and include six typical characteristics of wheat pests and diseases: baifen (Bf, powdery mildew), chimei (Cm, fusarium head blight), heisui (Hs, smut disease), yeman (Ym, wheat mite disease), qianying (Qy, leaf miner disease), and yachong (Yc, aphid disease). The shapes of these six characteristic features are illustrated in Fig. 6. These pests and diseases are all common in wheat, and training models capable of recognizing these pest and disease characteristics is of significant importance for promotion on farms.

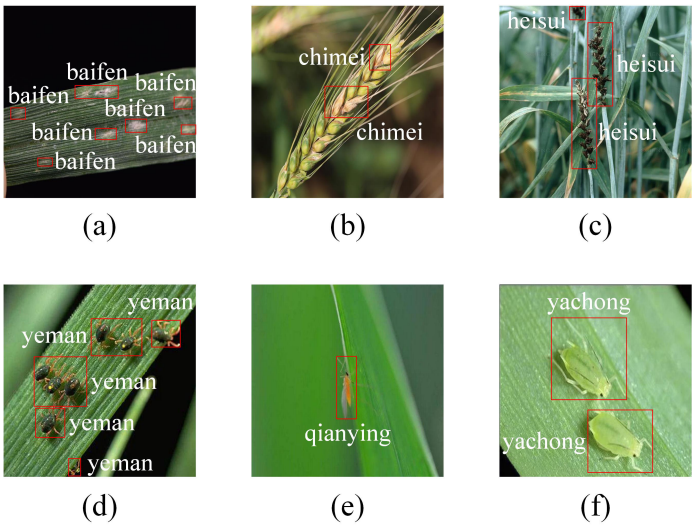

Fig. 6. Annotated representative features in Wheat-Data: (a) baifen, (b) chimei, (c) heisui, (d) yeman, (e) qianying, and (f) yachong.

In the experimental setup of this study, the operating system used is Linux, with an i9-14900HX CPU and an NVIDIA GeForce RTX 4090 GPU. The experiments are conducted using the PyTorch-2.4.0 deep learning framework, with CUDA-12.4 utilized for training acceleration.

B. Assessment Criteria

To comprehensively evaluate the model's accuracy, this study employs classic validation metrics such as precision, recall, average precision (AP), and mean average precision (mAP), with mAP as the primary evaluation metric. As shown in Eq. (13) to (16), the definitions of these metrics are as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{13}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{14}$$

$$AP = \int_0^1 p(r)dr \tag{15}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{16}$$

In classification evaluation metrics, the fundamental components are defined as follows: True Positives (TP) denote correct positive predictions, False Positives (FP) indicate erroneous positive classifications, and False Negatives (FN) represent undetected positive instances. The precision-recall relationship is mathematically characterized by the function $p(r)$, where n signifies the sample quantity within the i-th category. The detection performance for individual classes is quantified through Average Precision (AP), with AP_i specifically denoting the computed average precision for the i-th category.

To assess the model's computational efficiency and real-time capabilities, this study employs the Frames Per Second (FPS) metric as a key performance indicator. Specifically, FPS quantifies the maximum throughput achievable by the system by measuring how many image frames can be processed consecutively within one second. From an agricultural application perspective, higher FPS values directly correlate with enhanced real-time detection capacity for wheat pathogen symptoms and pest manifestations, which is particularly crucial for field deployment scenarios requiring instant diagnosis.

C. Results and Discussion

1) Comparative Analysis of Model Performance: To systematically assess the effectiveness of the proposed MGT-YOLO framework, this investigation conducts a comparative evaluation between the baseline YOLOv8 architecture and our enhanced MGT-YOLO implementation using the WheatData benchmark dataset. The quantitative evaluation results, including critical performance metrics of precision (P), recall (R), and mean average precision (mAP@0.5), have been comprehensively compiled in Table I for comparative analysis.

The comparative analysis presented in Table I reveals substantial performance enhancements achieved by the MGT-YOLO detection framework on the WheatData benchmark. Our architecture demonstrates a 2.0% absolute improvement in mean average precision (mAP@0.5) over the baseline YOLOv8 implementation, accompanied by consistent precision (P) gains across all feature categories. Particularly noteworthy are the 5.3% and 6.8% relative mAP@0.5 increments observed for the Bf and Hs detection tasks, respectively. These quantitative metrics substantiate the framework's superior efficacy in precisely identifying phytopathological characteristics associated with wheat crop infestations.

TABLE I. DETECTION PERFORMANCE OF YOLOV8 AND MGT-YOLO ON WHEATDATA

Dataset	Methods	Detect Type	P%	R%	mAP%
Wheat-Data	YOLOv8	Bf	75.1	80.6	84.7
		Ch	81.2	81.5	85.0
		Hs	62.4	50.7	58.7
		Ym	94.7	98.9	98.7
		Qy	95.9	99.6	99.2
		Yc	89.8	99.2	98.7
	MGT-YOLO	Bf	83.4	82.2	90.0(↑5.3)
		Ch	81.6	79.6	84.7(↓0.3)
		Hs	68.2	51.8	65.5(↑6.8)
		Ym	94.9	99.5	98.8(↑0.1)
		Qy	97.0	99.3	99.4(↑0.2)
		Yc	89.8	99.2	98.5(↓0.2)

The integration of our novel Multi-scale Enhancement Attention Mechanism (MEAM) into the backbone network architecture significantly augments feature extraction capabilities. Through systematic architectural innovation, the redesigned C2f_GlobalContext module in the neck network incorporates global context-aware operators that explicitly model cross-regional contextual dependencies, thereby effectively capturing long-range spatial-semantic relationships within agronomic visual data.

A comparative analysis was conducted to evaluate the small-object detection performance between the proposed MGT-YOLO framework and the baseline YOLOv8 model. We constructed precision-recall (PR) curves from experimental data. The PR curves for both methods are shown in Fig. 7, where Fig. 7(a) represents the PR curve of the baseline model, and Fig. 7(b) represents the PR curve of the MGT-YOLO model.The results demonstrate that MGT-YOLO achieves a larger area under the PR curve compared to its baseline counterpart.The overall mAP@0.5 achieved by the MGT-YOLO model on the WheatData dataset is 89.5%, surpassing the baseline model by 2.0 percentage points. It is worth noting that the Bf, Hs, and Ch features primarily appear as small objects. From the PR curve plots, it can be observed that the PR curve area for detecting these three small-object features is significantly larger for the MGT-YOLO model compared to YOLOv8.

To evaluate the performance of the MGT-YOLO model for each feature in the WheatData dataset, the study compares the mAP@0.5 values of several classical models on wheat pest and disease features within this dataset. The results are presented in Tables II and III.

Compared to other models, the MGT-YOLO model demonstrates superior overall detection accuracy as well as the best accuracy for each individual feature. This advantage is particularly evident for the Bf, Hs, and Ch wheat disease features, which are characterized by their small-object distribution. The enhanced performance can be attributed to the integration of the MEAM module, which further extracts high-level features of wheat diseases, and the C2f_GlobalContext module in the

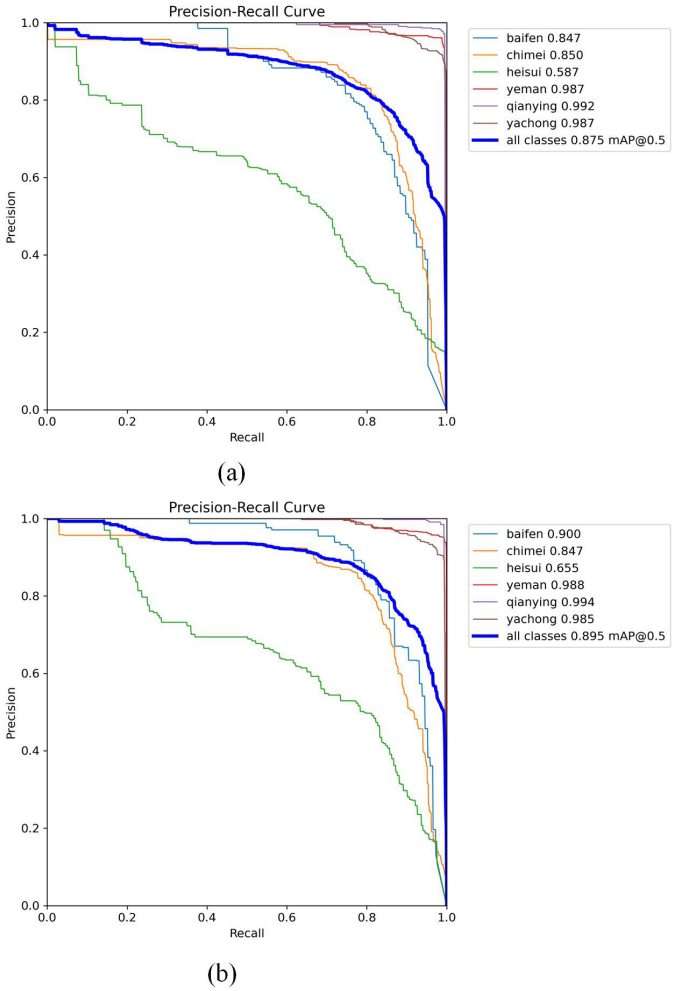

Fig. 7. The PR curves for YOLOv8 and MGT-YOLO on the WheatData. (a) The PR curves of YOLOv8 on the WheatData dataset. (b) The PR curves of MGT-YOLO on the WheatData dataset.

TABLE II. THE DETECTION RESULTS ON WHEAT-DATA DATASET

Dataset	Methods	mAP@0.5/%	GFLOPS	Parameters	FPS
Wheat-Data	Faster R-CNN	84.2	83.4	51.3 M	34.0
	SSD	83.4	30.6	34.5 M	53.3
	Transformer	85.2	126.2	50.5 M	46
	RetinaNet [36]	73.2	74.5	46.4 M	38.7
	YOLOX [37]	80.3	26.8	9.9 M	79.2
	YOLOv7 [38]	85.9	103.2	46.5 M	50.7
	YOLOv7-tiny [38]	82.6	13.1	7.0 M	96.2
	YOLOv8	87.5	8.1	4.6 M	179.2
	MSC-DNet [39]	88.1	78.6	44.1 M	90.0
	BHC-YOLO [27]	88.3	9.6	10.6 M	140.5
	MGT-YOLO	**89.5**	**9.2**	**5.9 M**	**161.4**

neck network. By capturing long-range dependencies in visual data, the C2f_GlobalContext module achieves stronger feature flow and facilitates more effective feature fusion.

As evidenced by the quantitative benchmarking in Table III, we conducted a visual comparative analysis between MGT-YOLO and selected single-stage detection networks that demonstrated optimal trade-offs in overall accuracy, model

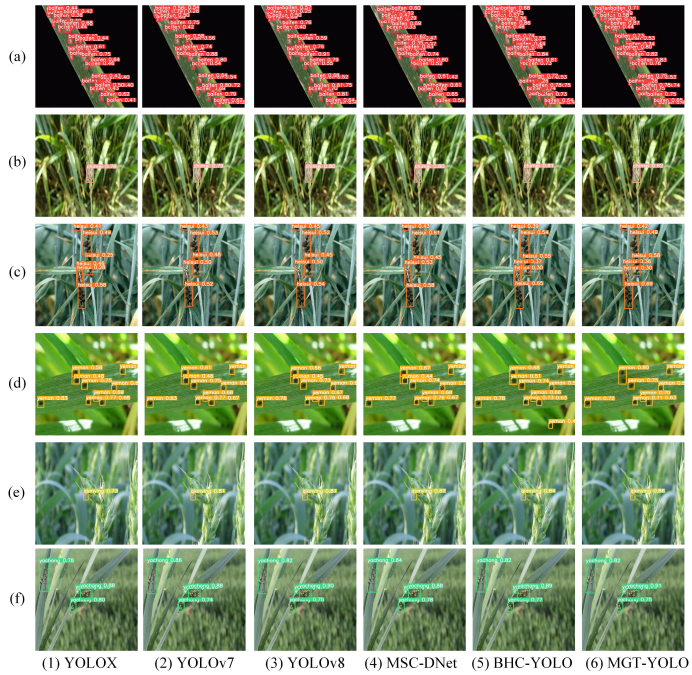

Fig. 8. Performance of MGT-YOLO and other comparative models on the Wheat-Data.

TABLE III. DETECTION RESULTS FOR EACH TYPE ON THE WHEATDATA

Types	YOLOv8	Faster R-CNN	RDD-YOLO	DsP-YOLO	MGT-YOLO
Baifen	84.7	83.6	85.2	85.0	**90.0**
Chimei	85.0	79.3	84.6	**85.2**	84.7
Heisui	58.7	57.6	55.5	58.5	**65.5**
Yeman	98.7	95.6	98.6	98.7	**98.8**
Qianying	99.2	95.3	98.5	98.9	**99.4**
Yachong	98.7	94.3	96.3	**98.7**	98.5
Overall mAP	87.5	84.2	88.1	88.3	**89.5**

compactness, and inference efficiency suitable for edge computing deployment. Fig. 8 provides a comprehensive visualization of detection outcomes across these models on the WheatData dataset. As depicted in the comparative results, MGT-YOLO exhibits markedly superior performance in capturing fine-grained pest and disease characteristics, particularly demonstrating enhanced detection precision for small-scale pathological features when contrasted with benchmark models.

As shown in Fig. 9, the confusion matrix of the proposed MGT-YOLO on the WheatData dataset is presented. From the analysis of Fig. 9, it can be observed that the model performs well in most categories, especially with minimal misclassification between the “qianying” and “wenku” classes. However, there is significant confusion between the “heisui” and “chimei” classes, with a relatively high misclassification rate between the two. This is due to the fact that both diseases occur in the spike part of the wheat. Despite this, the MGT-YOLO framework shows significant improvement compared to the baseline model. The misclassification rate in other categories is low, demonstrating good recognition capabilities.

As shown in Fig. 10, this is the performance result of MGT-YOLO on the WheatData dataset. In the figure, the top-left corner displays a bar chart where the x-axis represents

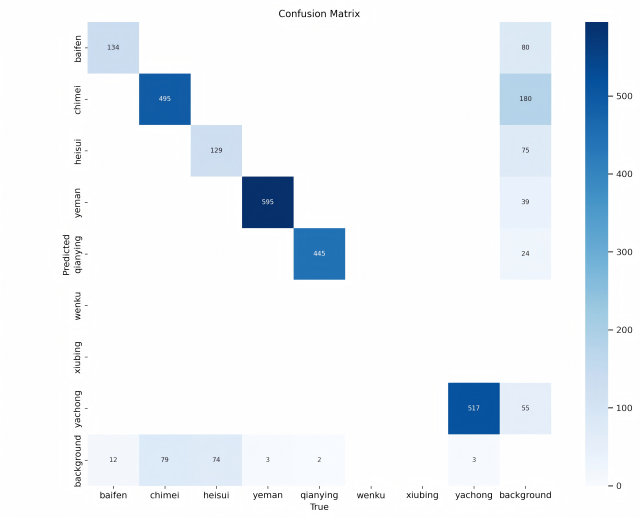

Fig. 9. Confusion matrix of the proposed MGT-YOLO on the WheatData dataset.

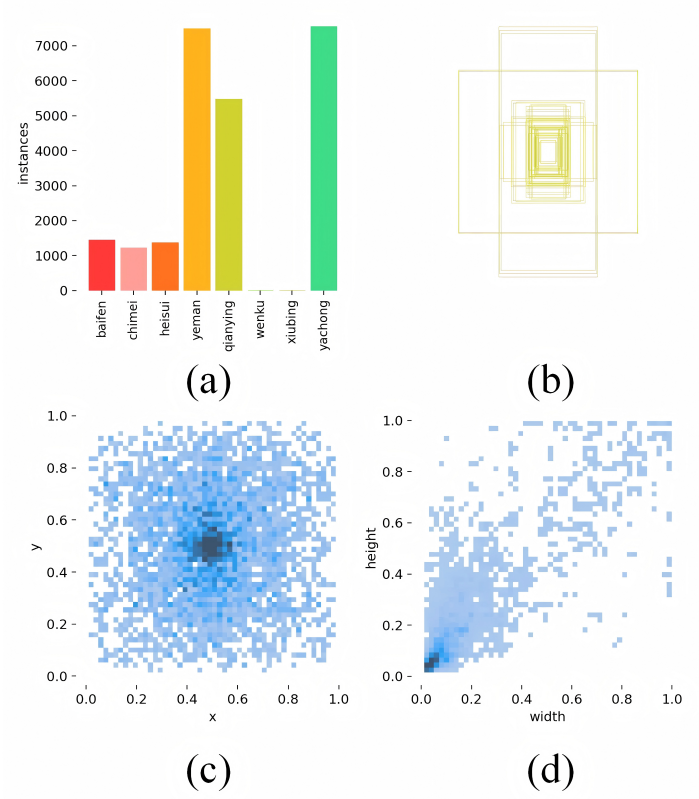

Fig. 10. Data analysis of the proposed MGT-YOLO on the Wheat-Data dataset. (a) Number of instances for different categories; (b) Distribution of bounding box sizes for different object categories; (c) Distribution of the center points of bounding boxes in the image; (d) Relationship between the height and width of the bounding boxes.

different categories such as baifen, chimei, heisui, and the y-axis represents the number of instances for each category. From the chart, it can be seen that the ”yeman” and ”yachong” categories have the highest number of instances, while ”baifen,” ”chimei,” and ”heisui” have relatively fewer instances. The box plot in the top-right corner shows the distribution of bounding box sizes for the MGT-YOLO model across different target categories. The scatter plot in the bottom-left corner illustrates the distribution of the x and y variables. The x-axis represents the position of the center of the bounding box along the image width, usually normalized with a range of [0, 1], while the y-axis represents the position of the center of the bounding box along the image height, also typically normalized with a range of [0, 1]. This plot demonstrates the distribution of the centers of different detection boxes in the image. The denser the scatter points, the more concentrated the bounding boxes are in that area. This plot shows that the data points are clustered around the central region, where the targets tend to appear more frequently. The scatter plot in the bottom-right corner shows the relationship between the height and width of the bounding boxes. The points are scattered, indicating a certain correlation between the height and width variables, with the density of points being mainly concentrated towards the lower part of the graph.

Fig. 11 shows the relationship between the center position and size of the MGT-YOLO detection boxes. Here, x and y represent the normalized coordinates of the center of the target box, and the histogram shows that the centers of the target boxes are generally concentrated in the central region of the image. Width and height represent the normalized width and height of the target boxes, and their distribution is more dispersed, indicating significant variation in the size of the target boxes. The scatter plot demonstrates the correlation between the variables, with x and y showing a certain concentration trend, while width and height exhibit a strong positive correlation.

2) Ablation Study: To systematically evaluate the individual and combined contributions of the MEAM, C2f_GlobalContext, and Vision Transformer modules to model performance, we performed a series of ablation studies on the WheatData dataset. As detailed in Table IV, the baseline YOLOv8n architecture achieves a mean average precision of 87.5% at IoU threshold 0.5 (mAP@0.5), while maintaining computational efficiency with 5.0M parameters and sustaining real-time inference speed at 179.2 frames per second. This experimental framework establishes a quantitative foundation for assessing the incremental improvements brought by each architectural enhancement.

TABLE IV. THE ABLATION EXPERIMENTS ON WHEATDATA

Methods	mAP@0.5/%	Parameters	FPS
Baseline	87.5	5.0 M	179.2
+ MEAM	88.3	5.5 M	153.2
+ C2f_GlobalContext	88.8	5.3 M	153.5
+ Vision Transformer	88.3	**4.9 M**	**181.0**
+ MEAM + C2f_GlobalContext	89.0	5.6 M	178.8
+ MEAM + Vision Transformer	88.3	5.3 M	155.1
+ C2f_GlobalContext + Vision Transformer	88.8	5.7 M	177.6
MGT-YOLO	**89.5**	5.9 M	161.4

The integration of the MEAM module into the backbone network demonstrates a 0.8% improvement in mAP@0.5, accompanied by a moderate computational cost increase of 0.5M parameters and a marginal reduction in inference speed (26.0 FPS decrease). Building on this, by integrating the

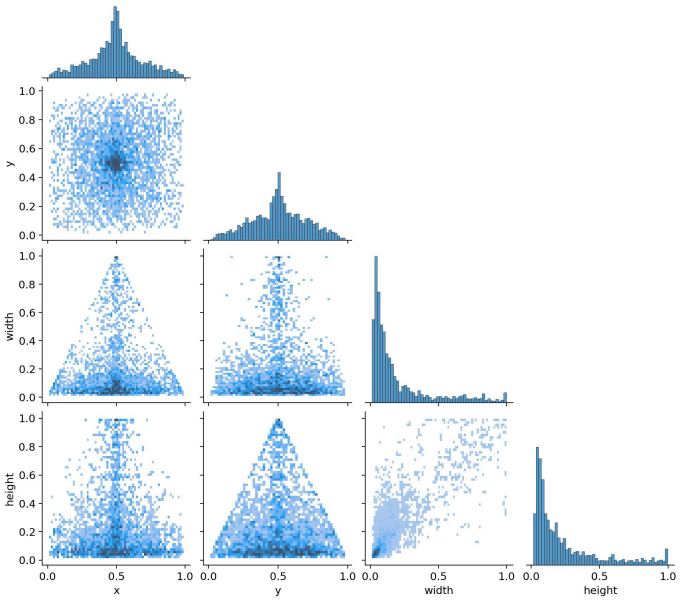

Fig. 11. Relationship between the center position and size of the MGT-YOLO detection boxes.

C2f_GlobalContext module into the neck network to capture long-range visual dependencies, we observe a further 0.7% enhancement in mAP@0.5, achieving the great performance of 89.0%. This modification slightly increases the parameter count and reduces FPS further. With the introduction of the Vision Transformer, the mAP@0.5 reaches 89.5%. Although the parameter count increases further, the real-time detection requirement is still met.

Compared to other models in the ablation experiments, the MGT-YOLO model achieves the greatest performance while maintaining an FPS comparable to the baseline, making it suitable for real-time wheat pest and disease detection tasks despite the added parameters.

V. Conclusion

This paper proposes the MGT-YOLO network, which integrates a multi-scale edge enhancement mechanism and a visual long-range dependency capturing mechanism to address the challenges of small-object recognition in wheat pest and disease detection under complex backgrounds. By introducing the Multi-scale Edge Enhancement Mechanism (MEAM), the Global Context Feature Fusion Module (C2f_GlobalContext), and incorporating the Vision Transformer module, the model significantly improves its ability to extract and integrate small-object pest and disease features. Experimental results on the self-constructed WheatData dataset demonstrate that MGT-YOLO outperforms traditional methods in detecting powdery mildew and smut, achieving an overall mAP@0.5 of 89.5%, significantly surpassing methods from related studies. The research shows that MGT-YOLO not only excels in small-object pest and disease detection but also holds potential for real-time applications in agricultural pest and disease management, providing crucial support for intelligent agricultural detection technologies.

The improvements in this paper are based on the YOLOv8 algorithm, focusing on feature extraction and feature fusion. This algorithmic model requires data collection and training for typical features, lacking universality in detection tasks. In the future, we plan to explore universal agricultural pest and disease detection tasks by incorporating multimodal large models.

Authors' Contributions

Conceptualization, Dandan Zhong and Penglin Wang; methodology, Dandan Zhong and Penglin Wang; validation, Dandan Zhong; formal analysis, Dandan Zhong; investigation, Dandan Zhong; resources, Jie Shen and Dongxu Zhang; writing—original draft preparation, Dandan Zhong; writing—review and editing, Jie Shen and Dongxu Zhang; visualization, Dandan Zhong; supervision, Penglin Wang, Jie Shen and Dongxu Zhang; project administration, Dandan Zhong; funding acquisition, Dongxu Zhang. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

This work was supported by the Central Government's Guide to Local Science and Technology Development Fund (YDZJSX2022C015), the Shanxi Provincial Basic Research Program (202303021221100), the Shanxi Agricultural University Science and Technology Innovation Enhancement Project (CXGC2023063), and the Shanxi Provincial Modern Agricultural Industry Technology System Construction Special Fund (2024CYJSTX02-14).

References

[1] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, "Automated visual defect detection for flat steel surface: A survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 3, pp. 626–644, 2020.

[2] Y. Zhang, H. Zhang, Q. Huang, Y. Han, and M. Zhao, "Dsp-yolo: An anchor-free network with dspan for small object detection of multiscale defects," *Expert Systems with Applications*, vol. 241, pp. 122 669–122 685, 2024.

[3] X. Dong, C. Zhang, J. Wang, Y. Chen, and D. Wang, "Real-time detection of surface cracking defects for large-sized stamped parts," *Computers in Industry*, vol. 159, pp. 104 105–104 119, 2024.

[4] Y. Gao, L. Gao, X. Li, and X. Yan, "A semi-supervised convolutional neural network-based method for steel surface defect recognition," *Robotics and Computer-Integrated Manufacturing*, vol. 61, pp. 101 825–101 832, 2020.

[5] R. Wang, H. Yu, J. Tang, B. Feng, Y. Kang, and K. Song, "Optimal design of iron-cored coil sensor in magnetic flux leakage detection of thick-walled steel pipe," *Measurement Science and Technology*, vol. 34, no. 8, pp. 085 123–085 133, 2023.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[7] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "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, 2016, pp. 21–37.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[11] A. A. Goudah, M. Jarofka, M. El-Habrouk, D. Schramm, and Y. G. Dessouky, "Object detection in inland vessels using combined trained and pretrained models of yolo8." *Advances in Computing & Engineering*, vol. 3, no. 2, pp. 751–766, 2023.

[12] R. Tian and M. Jia, "Dcc-centernet: A rapid detection method for steel surface defects," *Measurement*, vol. 187, pp. 110 211–110 225, 2022.

[13] Y. Wang, K. Zhang, L. Wang, and L. Wu, "An improved yolov8 algorithm for rail surface defect detection," *IEEE Access*, vol. 3, no. 2, pp. 751–766, 2024.

[14] C. Wang, H. Wang, Q. Han, Z. Zhang, D. Kong, and X. Zou, "Strawberry detection and ripeness classification using yolov8+ model and image processing method," *Agriculture*, vol. 14, no. 5, pp. 751–770, 2024.

[15] A. Ghafar, C. Chen, S. A. A. Shah, Z. U. Rehman, and G. Rahman, "Visualizing plant disease distribution and evaluating model performance for deep learning classification with yolov8," *Pathogens*, vol. 13, no. 12, pp. 1032–1047, 2024.

[16] J. Lin, G. Hu, and J. Chen, "Mixed data augmentation and osprey search strategy for enhancing yolo in tomato disease, pest, and weed detection," *Expert Systems with Applications*, vol. 264, pp. 125 737–125 752, 2025.

[17] L. Jia, T. Wang, Y. Chen, Y. Zang, X. Li, H. Shi, and L. Gao, "Mobilenet-ca-yolo: An improved yolov7 based on the mobilenetv3 and attention mechanism for rice pests and diseases detection," *Agriculture*, vol. 13, no. 7, pp. 1285–1300, 2023.

[18] J. Deng, C. Yang, K. Huang, L. Lei, J. Ye, W. Zeng, J. Zhang, Y. Lan, and Y. Zhang, "Deep-learning-based rice disease and insect pest detection on a mobile phone," *Agronomy*, vol. 13, no. 8, pp. 2139–2154, 2023.

[19] Y. Wang, R. Xu, D. Bai, and H. Lin, "Integrated learning-based pest and disease detection method for tea leaves," *Forests*, vol. 14, no. 5, pp. 1012–1027, 2023.

[20] J. Wang, J. Gao, and B. Zhang, "A small object detection model in aerial images based on cpdd-yolov8," *Scientific Reports*, vol. 15, no. 1, p. 770, 2025.

[21] Y. Zhang, G. Gao, Y. Chen, and Z. Yang, "Odd-yolov8: an algorithm for small object detection in uav imagery," *The Journal of Supercomputing*, vol. 81, no. 1, pp. 1–17, 2025.

[22] J. Qu, Q. Li, J. Pan, M. Sun, X. Lu, Y. Zhou, and H. Zhu, "Ss-yolov8: small-size object detection algorithm based on improved yolov8 for uav imagery," *Multimedia Systems*, vol. 31, no. 1, pp. 1–17, 2025.

[23] X. Zheng, J. Bi, K. Li, G. Zhang, and P. Jiang, "Smn-yolo: Lightweight yolov8-based model for small object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2025.

[24] W. Luo and S. Yuan, "Enhanced yolov8 for small-object detection in multiscale uav imagery: Innovations in detection accuracy and efficiency," *Digital Signal Processing*, vol. 158, p. 104964, 2025.

[25] P. Wang, D. Shi, and J. Aguilar, "Pcp-yolo: an approach integrating non-deep feature enhancement module and polarized self-attention for small object detection of multiscale defects," *Signal, Image and Video Processing*, vol. 19, no. 1, pp. 1–13, 2025.

[26] Z. Zhang, Y. Yang, X. Xu, L. Liu, J. Yue, R. Ding, Y. Lu, J. Liu, and H. Qiao, "Gvc-yolo: A lightweight real-time detection method for cotton aphid-damaged leaves based on edge computing," *Remote Sensing*, vol. 16, no. 16, pp. 3046–3061, 2024.

[27] B. Zhan, X. Xiong, X. Li, and W. Luo, "Bhc-yolov8: improved yolov8-based bhc target detection model for tea leaf disease and defect in real-world scenarios," *Frontiers in Plant Science*, vol. 15, pp. 1 492 504–1 492 519, 2024.

[28] H. Dong, M. Yuan, S. Wang, L. Zhang, W. Bao, Y. Liu, and Q. Hu, "Pham-yolo: A parallel hybrid attention mechanism network for defect detection of meter in substation," *Sensors*, vol. 23, no. 13, pp. 6052–6061, 2023.

[29] J. Wang and J. Wang, "A lightweight yolov8 based on attention mechanism for mango pest and disease detection," *Journal of real-time image processing*, vol. 21, no. 4, pp. 136–151, 2024.

[30] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[31] G. Chen, Y. Hou, T. Cui, H. Li, F. Shangguan, and L. Cao, "Yolov8-cml: A lightweight target detection method for color-changing melon ripening in intelligent agriculture," *Scientific Reports*, vol. 14, no. 1, pp. 14 400–14 410, 2024.

[32] S. Gao, P. Zhang, T. Yan, and H. Lu, "Multi-scale and detail-enhanced segment anything model for salient object detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9894–9903.

[33] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 1–15.

[34] S. Yun and Y. Ro, "Shvit: Single-head vision transformer with memory efficient macro design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5756–5767.

[35] S. Du, B. Zhang, P. Zhang, and P. Xiang, "An improved bounding box regression loss function based on ciou loss for multi-scale object detection," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*. IEEE, 2021, pp. 92–98.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[37] T. Panboonyuen, S. Thongbai, W. Wongweeranimit, P. Santitamnont, K. Suphan, and C. Charoenphon, "Object detection of road assets using transformer-based yolox with feature pyramid decoder on thai highway panorama," *Information*, vol. 13, no. 1, pp. 5–15, 2021.

[38] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "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*, 2023, pp. 7464–7475.

[39] R. Liu, M. Huang, Z. Gao, Z. Cao, and P. Cao, "Msc-dnet: An efficient detector with multi-scale context for defect detection on strip steel surface," *Measurement*, vol. 209, pp. 112 467–112 482, 2023.