The Visual Computer (2026) 42:61
https://doi.org/10.1007/s00371-025-04225-9

RESEARCH

HSFPN-Det: an effective model for detecting rice pests and diseases

Yang Yang[1,2] · Yuxin Hong[1,2] · Wenjie Yu[1,2] · Xiao Zhang[1,2,3] · Bo Yang[1,2] · Meng Shi[1,2] · Yangguang Sun[1,2] · Jun Wang[1,2] · Jianlin Zhu[1,2]

Received: 9 May 2025 / Accepted: 29 October 2025

Abstract

Rice pests and diseases pose a significant challenge to global food security, which is exacerbated by the increasing frequency of climate change and extreme weather events. Conventional manual techniques for the identification of rice pests and diseases are characterized by inefficiency and limited precision. Additionally, they are heavily dependent on pre-existing expertise, rendering them inadequate for application in large-scale detection contexts. In addressing these challenges, this study presents HSFPN-Det, a novel lightweight model developed for the detection of rice pests and diseases. The proposed model integrates a High-level Selective Feature Pyramid Network (HSFPN) with a deformable self-attention mechanism and a Hybrid Attention Transformer, with the objective of improving both detection effectiveness and accuracy. Furthermore, by employing an enhanced network architecture, the neck network is optimized to achieve better multi-scale feature fusion. The deformable self-attention module has been developed to dynamically concentrate on prominent spatial features. The model is evaluated on an augmented rice pest dataset containing 11,292 images across four rice disease categories. In comparative analyses with cutting-edge models, including YOLOv11, YOLOv9, SSD, and EfficientDet-D1, it consistently yields superior performance outcomes. Notably, its compact model size of 3.97 MB is approximately 49.87% smaller than previous models. Additionally, further evaluations conducted on real-world images confirm its superior detection accuracy in diverse scenarios. HSFPN-Det achieves higher accuracy while reducing model complexity, offering a practical and deployable solution for smart agriculture. Its ability to run effectively on low-resource devices marks a crucial contribution to real-time intelligent pest management in precision agriculture.

Keywords Rice pests and diseases detection · Deformable self-attention · Hybrid attention transformer · Selective feature fusion

1 Introduction

It is evident that as global climate anomalies become increasingly severe, there is a concomitant rise in the frequency of extreme meteorological disasters [1]. This change severely

✉ Jianlin Zhu
Jianlin.Zhu@mail.scuec.edu.cn

Yang Yang
yangyang@mail.scuec.edu.cn

Yuxin Hong
hongyuxin@ln.hk

Wenjie Yu
18949874102@163.com

Xiao Zhang
xiao.zhang@my.cityu.edu.hk

Bo Yang
yangbo@mail.scuec.edu.cn

Meng Shi
mengshi@mail.scuec.edu.cn

Yangguang Sun
ygsun@mail.scuec.edu.cn

Jun Wang
junwang@scuec.edu.cn

[1] School of Computer Science (School of Artificial Intelligence), South-Central Minzu University, No.182, Minyuan Road, Wuhan 430074, Hubei Province, China

[2] Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprise, South-Central Minzu University, No.182, Minyuan Road, Wuhan 430074, Hubei Province, China

[3] School of Computer Science and Technology, University of Science and Technology of China, No.96, JinZhai Road, Hefei 230026, Anhui Province, China

Published online: 14 December 2025

affected the normal growth of rice and significantly increased the outbreak of rice plant diseases and insect pests, which in turn led to a reduction in rice yield. In this context, the establishment of a rapid and accurate system for the identification of crop diseases and pests has become an urgent necessity. Previous identification methods primarily relied on the visual observation of samples by agricultural experts from plant protection agencies [2]. However, the efficacy of this manual identification approach is contingent upon the expertise of the inspectors, thus exposing evident limitations, including its time-consuming nature, inefficiency, and the presence of substantial subjective bias in the judgment outcomes.

Recent research has demonstrated notable progress in applying enhanced Swin Transformer architectures for object detection [3, 4]. The Transformer's attention mechanism is particularly effective in processing images with blurred or complex backgrounds. Moreover, its hierarchical structure, combined with a shifted window strategy, significantly reduces computational complexity. This advancement permits Swin Transformer-based models to be implemented on embedded devices, thereby fulfilling the objective of real-time detection of rice pests and diseases.

In order to mitigate the challenges of diminished detection efficacy and the absence of immediate response capabilities in the identification of rice diseases and pests, this study presents an advanced detection framework, designated as HSFPN-Det. By integrating multi-scale feature information with a deformable adaptive attention mechanism [5], the framework deeply excavates the potential of deep learning algorithms. At the technical implementation level, the system first employs a Hybrid Attention transformer to perform high-resolution reconstruction on input rice image data, followed by robust feature extraction operations. Through the organic combination of channel attention strategy with deconvolution and conventional convolutional layers, effective feature fusion and enhancement are achieved. The optimized feature information is then fed into the detection module to ultimately accomplish accurate identification of disease and pest types. The core of HSFPN-Det's outstanding detection performance lies in two aspects: first, the adoption of an innovative feature fusion component to replace the traditional neck connection structure, which can effectively integrate deep and shallow feature representations; second, the introduction of a deformable adaptive attention unit that enables dynamic adjustment and weight allocation of input features, a design concept consistent with related research findings. The data used in the study originates from a collection of rice disease and pest images captured in controlled environments. Employing traditional data augmentation techniques, including image flipping, scale transformation, and region cropping, a robust training sample library was established. Experimental validation based on this dataset indicates that the proposed model attains industry-leading performance, exhibiting substantial advantages over existing comparative algorithms.

Our main contribution is the following:

- The proposed High-level Selective Feature Pyramid Network (HSFPN) constitutes a breakthrough advancement in this technical domain. This network architecture abandons conventional connection module designs and demonstrates superior processing capabilities in integrating deep and shallow feature information. Experimental validation confirms that HSFPN can achieve significant improvements in detection accuracy for object recognition tasks across different scales.
- A feature enhancement module is introduced by integrating deformable self-attention mechanisms with Hybrid Attention transformers, which not only improves adaptability to variations in object shape and spatial position, but also strengthens image feature representation, thereby enhancing detection accuracy.
- Through rigorous comparative experimentation involving an array of prevalent models, such as YOLOv5, YOLOv8, SSD, and EfficientDet-D1, alongside meticulous ablation studies exploring diverse attention mechanisms and structural configurations, the proposed model has undergone thorough validation. This validation underscores its efficacy, lightweight architecture, and outstanding performance, thereby accentuating its significant promise for practical deployment in agricultural applications.

2 Related work

2.1 Rice disease and pest detection based on traditional methods

In the current development of object detection technology, researchers typically combine handcrafted features with traditional machine learning classification algorithms to accomplish object recognition tasks. In the field of rice pest and disease detection, Guo et al. [6]. enhanced the detection capability of small objects by introducing a lightweight CARAFE operator, addressing the challenges of feature extraction difficulties and low detection accuracy in rice pest and disease recognition. Experimental results demonstrate that this method improves accuracy, recall rate, and mean average precision by 1.9%, 2.9%, and 2.7%, respectively, compared to YOLOv8n. Addressing the problem of rice pest and disease detection, researchers like Peng et al. [7] constructed the DC-GhostNet lightweight network architecture. This network integrates depthwise separable convolution technology, feature integration strategies, and an adaptive k-value IECA attention module, aiming to improve detec-

tion accuracy. The Ghost component in the network utilizes channel rearrangement operations to achieve global feature optimization, demonstrating recognition accuracy surpassing baseline models on a dataset containing 15 categories of rice pests and diseases, with a processing time of only 13.1 milliseconds per image. Zheng et al. [8] proposed an efficient rice pest and disease detection method based on the YOLOv8 architecture, which can identify 16 types of rice pests and diseases. After extensive experimental validation, the method's mAP increased from 61.1% to 70.2%, and the $F1$ score also improved from 60.0% to 67.7%, reaching state-of-the-art performance.

2.2 Rice disease and pest detection based on CNNs

In the realm of deep learning methodologies, several researchers have undertaken comprehensive investigations into the detection of agricultural pests and diseases. Mainstream deep learning frameworks include SSD, YOLO, R-CNN, and Faster R-CNN. Related research indicates that compared to traditional manual identification methods, deep learning technology can achieve more excellent results in crop pest and disease recognition tasks. Researchers like Sun et al. [9] utilized the SSD deep learning object detection algorithm to train rice data and constructed a rice panicle detection model. This method achieved a mAP of 38.1% on the validation set, opening new possibilities for neural network research related to rice and providing scientific support for rice yield estimation, thereby helping to better guide rice cultivation practices. The research collective spearheaded by Li et al. [10] has engineered an object detection architecture of exceptional precision and robustness, founded upon an enhanced real-time detection transformer structure. This advanced model, termed the Intelligent Multi-scale Litchi Leaf Pest Detection Transformer (IMLL-DETR), epitomizes cutting-edge technological innovation. By integrating dynamic convolution techniques, the model adeptly apprehends pivotal global contextual information, thereby effectively circumventing the domination by local features and minimizing disturbances from complex backgrounds and inter-target relationships. Empirical findings corroborate that IMLL-DETR exhibits outstanding robustness in conjunction with high-precision capabilities.

3 Methodology

The framework diagram of the HSFPN-Det proposed in this paper is shown in Fig. 1. This framework employs an innovative selective feature pyramid network to replace conventional connection modules, aiming to achieve better integration of deep and shallow feature information, thereby significantly improving the recognition effectiveness of targets at different scales. Meanwhile, a feature enhancement component is designed by integrating deformable self-attention mechanisms with Hybrid Attention transformer technology. Experimental validation demonstrates that this component significantly enhances the model's adaptability to variations in geometric shapes and spatial positioning of targets, thereby improving the expressive richness of image features and culminating in an increased accuracy of detection.

3.1 HSFPN-Det

In contemporary applications, the predominant paradigm employed by deep learning methodologies for detecting rice pests and diseases is a single-step execution model, which is conducive to achieving rapid detection speeds. Nonetheless, in comparison with multi-step algorithms, such methodologies often exhibit deficiencies in accuracy. This research is motivated by the HSFPN method integrated into the MFDS-DETR framework, with the objective of enhancing the hierarchical feature fusion process and improving the structural design of the original Neck module, as elaborated in [5]. Within the YOLOv8 framework, the Neck component serves as a critical intermediary between the backbone network and the detection head, enabling the integration and processing of features to bolster detection performance and precision. The HSFPN-Det model developed in this paper (elucidated in Fig. 1) represents a sophisticated enhancement of the current YOLOv8 algorithm. The backbone network architecture integrates an adapted version of CSPDarkNet, wherein the traditional C3 module is replaced by a C2f configuration. Additionally, the kernel size of the initial convolutional layer is reduced from 6×6 to 3×3 to alleviate computational overhead. The neck network, HSFPN, employs 1×1 convolution operations to standardize three feature maps of varying resolutions to a consistent 256-channel dimension, subsequently augmenting feature representation capabilities through channel attention mechanisms and selective feature fusion components. The detection head segment incorporates a deformable self-attention component encompassing three attention heads, employing linear layers for output processing. The decoupled detection head structure independently addresses regression prediction and classification prediction tasks.

3.2 Overall architecture of our HSFPN

Figure 2 shows the HSFPN framework, which consists of two ingenious components: the Feature Selection Component and the Feature Integration Component. The Feature Selection Component leverages channel attention mechanisms in conjunction with dimension-matching techniques to enhance feature maps across multiple scales [5]. This enhancement is

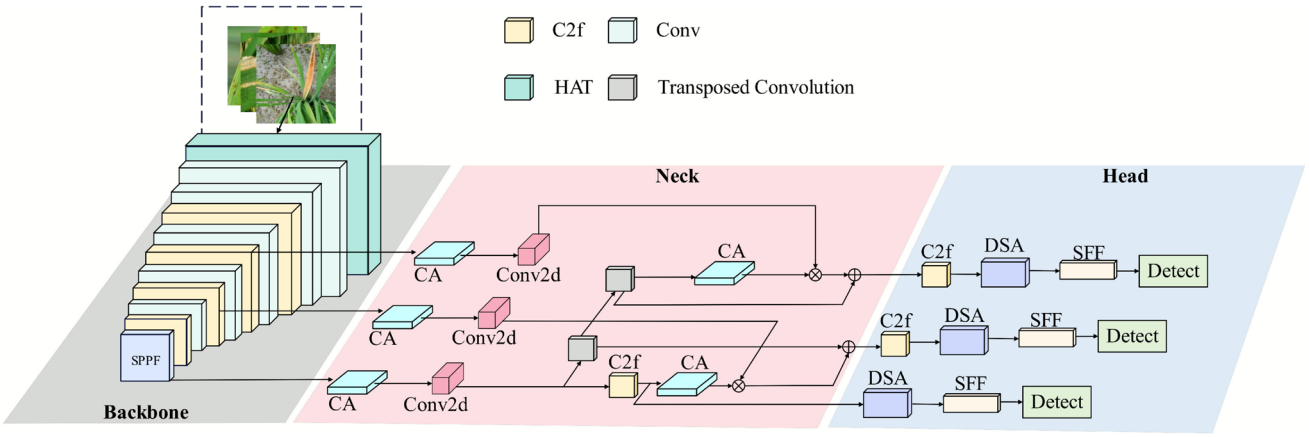

Fig. 1 The comprehensive architecture of the proposed HSFPN-Det framework, designed for the detection of rice pests and diseases, embodies a sophisticated approach. Our HSFPN primarily consists of three parts: Backbone, Neck, and Head. Initially, the input image is passed through the Backbone network, comprising layers such as SPPF, C2F, and Hybrid Attention Transformer, to extract fundamental features. These features are subsequently transmitted to the Neck module, where they undergo multi-path fusion and enhancement via the Channel Attention mechanism, convolutional layers, and transposed convolution. Finally, the enhanced feature representations are forwarded to the Head module, which incorporates components like C2F, DSA, and SFF, to produce the final detection results

achieved through global average and max pooling operations, which employ weight-based calculations to extract salient channel features. Simultaneously, the Feature Integration Component adopts a selective fusion approach to combine refined low-level features with corresponding high-level features. In this process, high-level features are expanded and rescaled using bilinear interpolation or deconvolution methods [11], enabling effective fusion. Collectively, these modules empower the HSFPN to adeptly address the challenges inherent in multi-scale detection, thereby substantially improving the accuracy and robustness of the detection system.

3.3 Overall architecture of the selective feature fusion module

The Selective Feature Fusion (SFF) module has been thoroughly validated to utilize high-level features as guiding weights, thereby effectively extracting critical semantic information from lower-level features, as evidenced in [5]. The architectural design of this advanced module is illustrated in Fig. 3. This methodical approach to feature fusion substantially enhances the module's performance and accuracy. Considering high-level features denoted by $f_{\text{high}} \in \mathbb{R}^{C\times H\times W}$ and low-level features denoted by $f_{\text{low}} \in \mathbb{R}^{C\times H_1\times W_1}$, a 3×3 convolution with a stride of 2 is applied to the high-level features, resulting in adjusted feature dimensions $f_{\widehat{\text{high}}} \in \mathbb{R}^{C\times 2H\times 2W}$. To reconcile the dimensional discrepancy between feature levels, a transposed convolution operation is subsequently employed to produce scaled high-level features $f_{\text{att}} \in \mathbb{R}^{C\times H_1\times W_1}$. Once dimensional alignment is achieved, the Channel Attention (CA) mechanism is activated. This sophisticated procedure culminates in the selective fusion of filtered low-level features with their corresponding high-level counterparts. This operation not only enhances the representational capacity of the model but also yields an output $f_{\text{out}} \in \mathbb{R}^{C\times H_1\times W_1}$. The subsequent equation formally describes the precise mechanism employed for the selection and integration of features:

$$f_{\text{att}} = BL(T - \text{Conv}(f_{\text{high}})) \tag{1}$$

$$f_{\text{out}} = f_{\text{low}} * CA(f_{\text{att}}) + f_{\text{att}} \tag{2}$$

3.4 The overall architecture of the deformable self-attention module

Figure 4 shows the deformable self-attention framework, the deformable self-attention framework comprises two main components: an offset computation module and an attention computation module [5]. The computational workflow initiates with the projection of input vectors into feature maps, which are subsequently processed by the offset module. Within this module, query vectors are formulated from reference point coordinates, followed by a linear transformation to compute positional offsets Δp_q. Concurrently, a parallel linear operation extracts content features from the input feature maps, with bilinear interpolation yielding the final offset values $\text{offset}_{\text{value}}$.

The attention module employs a dual transformation mechanism: first, a linear projection processes the query vector, after which the Softmax function generates normalized weight distributions for each offset [12]. These weights are

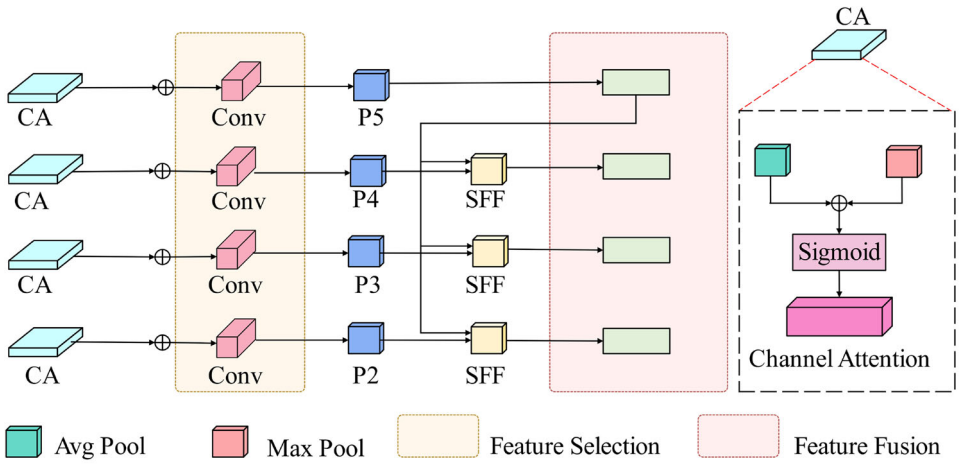

Fig. 2 The architecture of the HSFPN initiates with the implementation of a channel attention component, which allocates weights to the incoming data on a channel-specific basis. This initial step ensures that feature significance is appropriately calibrated. Following this, convolutional layers are employed for precise feature extraction, where the extracted features undergo refinement through a structured pyramid pooling hierarchy. The selected feature pathways are subsequently subjected to further optimization and enhancement within the selective feature fusion component. Ultimately, the feature integration module aggregates all the processed information, ensuring comprehensive consolidation. Within its intrinsic mechanism, the channel attention mechanism synergistically integrates average pooling and max pooling to ascertain spatial attention coefficients. These coefficients are subsequently activated via the sigmoid function, thereby accentuating the importance of salient feature channels

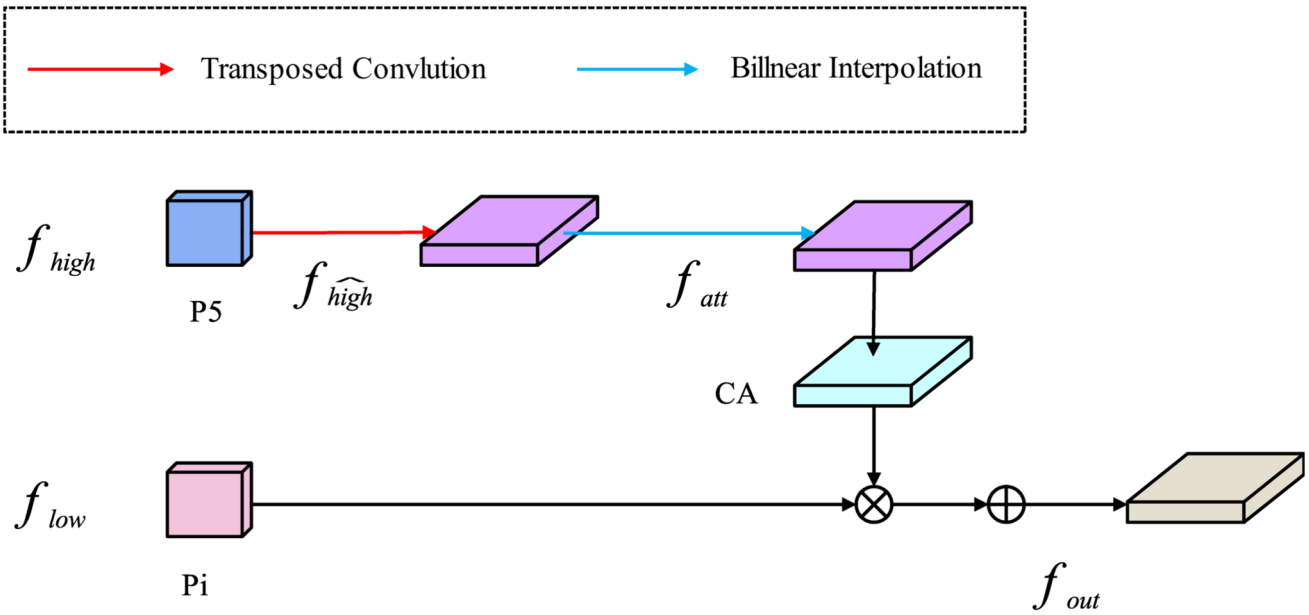

Fig. 3 The selective feature fusion module is designed as follows: the high-level feature f_{high} is first processed via the highest-level pooling layer (P5). Subsequently, the attention feature f_{att} is generated through a transposed convolution operation followed by bilinear interpolation. This attention feature is then refined using the Channel Attention (CA) module and multiplied element-wise with the low-level feature f_{low}, which is obtained from the Pi module, to perform feature weighting. Ultimately, the weighted feature is integrated with other feature maps to yield the final output f_{out}

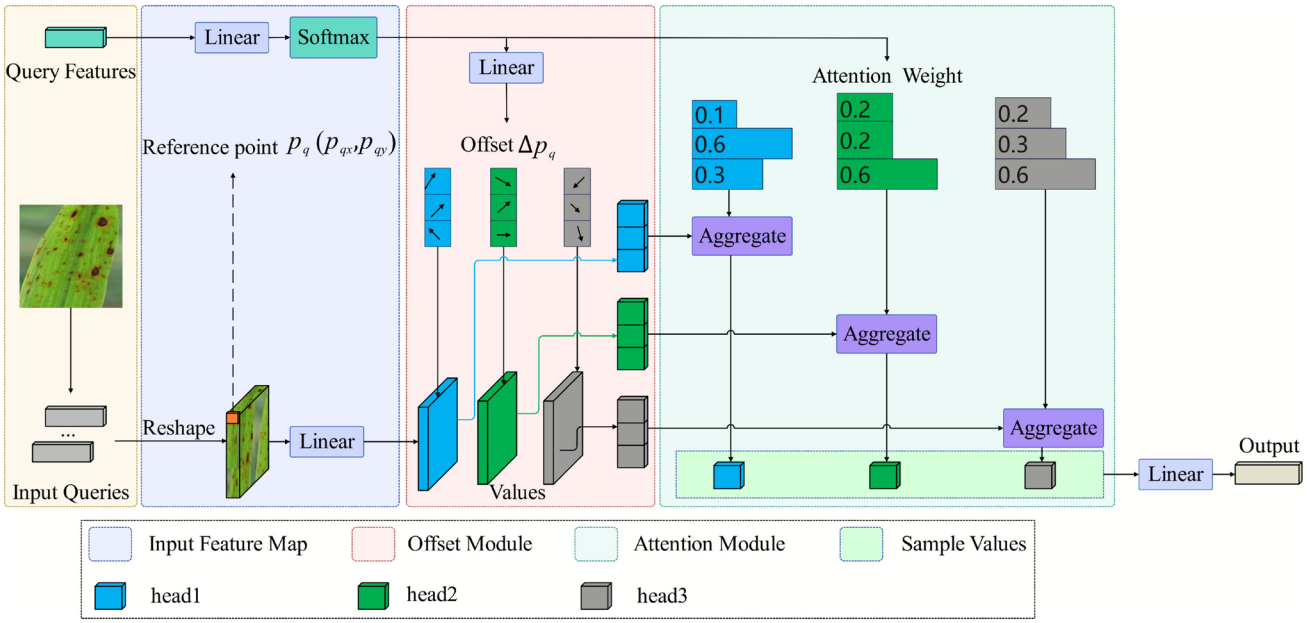

Fig. 4 The structural configuration of a deformable self-attention module is delineated as follows: The query features, presented on the left, are bifurcated into two simultaneous processing pathways. The upper trajectory comprises a linear transformation succeeded by a Softmax function to ascertain reference points p. In parallel, the lower pathway undergoes a reshaping operation followed by another linear transformation. Centrally positioned, an offset module is responsible for the derivation of positional offsets Δp. The resultant features are subsequently processed by three distinct attention heads, visually represented in blue, green, and gray hues, and amalgamated with trainable attention weights. These weighted features are synthesized through an aggregation module, and then channeled through a terminal linear layer to yield the final output. This comprehensive architecture employs a deformable multi-head self-attention mechanism, significantly augmenting the model's proficiency in encapsulating intricate spatial variations commonly observed in rice pest and disease environments

then combined with corresponding offset values through a weighted summation:

$$\text{Weight} = \text{Softmax}(WQ) \tag{3}$$

$$\text{Sample}_{\text{value}} = \sum_{k=1}^{K} \text{offset}_{\text{value}} \cdot \text{Weight} \tag{4}$$

Multi-head attention is realized through concatenation of outputs from H independent attention heads:

$$\text{Sample}_{\text{output}} = \text{Concat}(\text{Sample}^1_{\text{value}}, \ldots, \text{Sample}^H_{\text{value}}) \tag{5}$$

The final output is obtained via linear transformation of the aggregated features:

$$\text{Output} = W \cdot \text{Sample}_{\text{output}} \tag{6}$$

This hierarchical architecture enables adaptive feature aggregation, where learnable offsets dynamically adjust receptive fields while attention weights modulate feature importance [5]. The mathematical formulation preserves spatial relationships through bilinear interpolation while maintaining computational efficiency through separable linear operations.

Designed to concentrate computational resources on salient regions of input data, the Deformable Self-Attention Module transcends the constraints inherent to conventional attention frameworks, which indiscriminately process expansive contextual information. This mechanism demonstrates an intrinsic capacity to adapt its receptive fields in a context-dependent manner, thereby capturing a more comprehensive array of task-relevant feature representations. This inherent plasticity empowers the module to dynamically prioritize elements within the data that are critical both spatially and semantically. Consequently, the models exhibit heightened efficacy in deconstructing images marked by complex structural arrangements and significant morphological variability.

3.5 Hybrid attention transformer

As illustrated in Fig. 5, the Hybrid Attention Transformer architecture is presented with its complete structure and key components [13]. The framework comprises three principal modules: (1) Shallow Feature Extraction employing convo-

lutional operations, (2) The extraction of intricate features is facilitated through the deployment of multiple Residual Hybrid Attention Groups (RHAGs), which adeptly model both channel-wise and spatial interconnections throughout the feature extraction process, and (3) Image Reconstruction where convolutional operations restore high-resolution images from blurry inputs to enable precise identification of rice pests and diseases.

4 Experiment

To verify the effectiveness of our HSFPN-Det for rice pest and disease detection, this paper designs comparative experiments with other baseline models. The performance of HSFPN-Det was benchmarked against a range of prominent algorithms, including state-of-the-art detection techniques such as YOLOv5, YOLOv8, YOLOv9, YOLOv11, YOLO-World, SSD, EfficientDet-D1, TOOD, DINO, Deformable DETR, and DAB-DETR. The training regimen was set for 500 epochs to ensure adequate model learning. Throughout the training process, the loss function was continuously monitored in real time. A stabilization of the loss function across multiple consecutive epochs without a marked downward trend was interpreted as an indication that the model was nearing its optimal performance. At this juncture, prolonging the training could potentially lead to overfitting.

4.1 Implementation details

The empirical experiments were conducted using a single NVIDIA RTX 4090 GPU equipped with 120GB of memory and implemented within the PyTorch framework. During the training phase, the parameters were set as follows: 800 epochs, a batch size of 2, a learning rate of 1e−2, and a weight decay coefficient of 3e−5. The model's implementation relied on PyTorch version 2.1.0 for core functionalities, OpenCV version 4.12 for advanced image processing tasks, and TensorBoard version 2.20 for real-time visualization.

4.2 Dataset

In this study, a comprehensive dataset containing 11,292 images was systematically constructed and meticulously divided into three subsets: a training set containing 9,868 images, a validation set containing 949 images, and a test set containing 475 images; the dataset covers four major types of crop diseases, including 2,658 cases of leaf spot, 3,631 cases of brown spot, 1,979 cases of tungro disease, and 3,024 cases of bacterial wilt; to accurately reflect the complex conditions of real agricultural environments, the dataset was enhanced through various advanced data augmentation techniques, such as horizontal image flipping and light intensity adjustment, thereby improving its authenticity and robustness (Table 1).

Table 1 Dataset attribute parameters

Categories	Train	Validate	Test	Total
Leaf spot disease	2237	313	108	2658
Brownspot	3261	275	95	3631
Tungro	1692	163	124	1979
Bacterial blight	2678	198	148	3024

4.3 Comparisons with SOTA models

To evaluate the proficiency of the proposed algorithm within the realm of rice pest and disease identification, this investigation embarked upon a suite of thorough comparative experiments. A selection of cutting-edge detection algorithms—comprising YOLOv11, YOLOv9, YOLOv8, YOLOv5, YOLO-World, Single Shot MultiBox Detector (SSD), EfficientDet-D1, Task-Aligned Detection (TOOD), Detection Transformer with Improved Nonlinear Operations (DINO), Deformable DETR, and Deformable Attention-based DETR (DAB-DETR)–were employed for a robust performance evaluation. Moreover, results from ablation studies underscore that the HSFPN architecture introduced in this study markedly augments the feature fusion proficiency within the neck module of the network, surpassing the capabilities inherent in the original network framework.

The experimental outcomes reveal that the HSFPN-Det model, as proposed in this paper, excels in two critical evaluation metrics: the mean Average Precision at an IoU threshold of 0.5 (mAP@0.5) reaches a commendable value of 0.624, while the mean Average Precision across IoU thresholds (mAP@0.5:.95) attains a value of 0.303. This level of performance notably surpasses that of the diverse baseline models employed in the comparative experiments. According to the statistical data in Table 2, the five methods YOLOv11, TOOD, DINO, Deformable DETR, and DAB-DETR show relatively weak performance on the test dataset, with their mAP@0.5 scores being 0.221, 0.422, 0.266, 0.141, and 0.284, respectively, while their corresponding mAP@.5:.95 scores are 0.070, 0.230, 0.128, 0.053, and 0.159, respectively. The main reason for the poor performance of these algorithms may lie in the limitations in extracting key feature information from pest and disease images. In detail, the DINO algorithm draws inspiration from the learning strategy of word vector embeddings in the natural language processing field, while Deformable DETR improves the original DETR architecture by fusing multi-scale feature representations and multi-scale positional encoding mechanisms. Particularly noteworthy is that the YOLO-World model has a storage

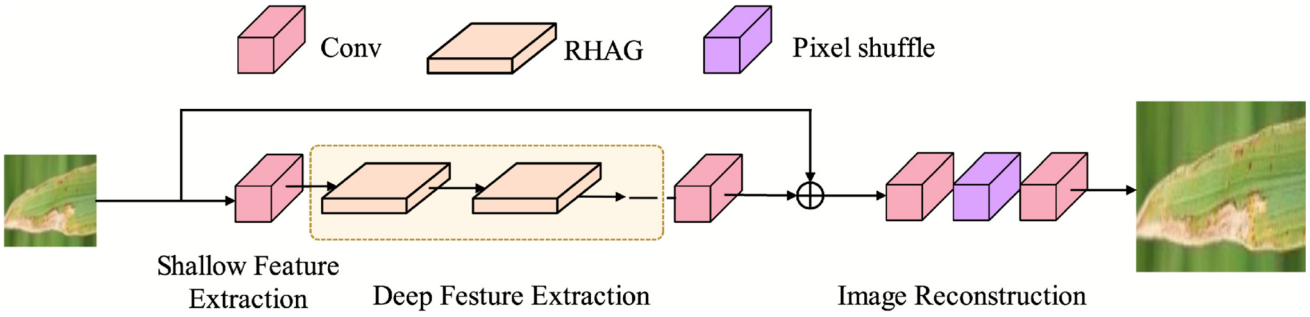

Fig. 5 Structural diagram of Hybrid Attention Transformer featuring residual channel-spatial attention mechanisms. The input image is first processed by a shallow feature extraction module to obtain basic feature representations. These features are subsequently forwarded to the deep feature extraction stage, which comprises multiple cascaded Residual Hybrid Attention Groups(RHAG). The features extracted from this stage are then combined with the original shallow features via a residual connection. Finally, the reconstructed high-resolution image is generated through the image reconstruction module

Table 2 Comparison tests of different models, the final row illustrates the experimental outcomes of the HSFPN-Det model applied to the PlantVillage dataset. [Key: **Best**, *Second Best*]

Methods	mAP@0.5	mAP@.5:.95	Recall	Model weights
YOLOv11	$0.221_{\pm 0.103}$	$0.070_{\pm 0.062}$	$0.384_{\pm 0.327}$	*5.25MB*
YOLOv9	$0.604_{\pm 0.071}$	$\mathit{0.294}_{\pm \mathit{0.105}}$	$0.626_{\pm 0.212}$	6.27MB
YOLOv8	$\mathit{0.605}_{\pm \mathit{0.023}}$	$0.291_{\pm 0.019}$	$0.623_{\pm 0.057}$	5.95MB
YOLOv5	$0.589_{\pm 0.301}$	$0.260_{\pm 0.226}$	$\mathbf{0.658}_{\pm \mathbf{0.239}}$	13.70MB
YOLO-World [14]	$0.474_{\pm 0.426}$	$0.223_{\pm 0.551}$	$0.459_{\pm 0.637}$	520.00MB
SSD [15]	$0.420_{\pm 0.228}$	$0.120_{\pm 0.217}$	$0.360_{\pm 0.173}$	92.10MB
EfficientDet-D1 [16]	$0.324_{\pm 0.214}$	$0.261_{\pm 0.328}$	$0.351_{\pm 0.335}$	25.60MB
TOOD [17]	$0.422_{\pm 0.736}$	$0.230_{\pm 0.655}$	$0.471_{\pm 0.604}$	246.00MB
DINO [18]	$0.266_{\pm 0.509}$	$0.128_{\pm 0.631}$	$0.529_{\pm 0.715}$	407.00MB
Deformable DETR [19]	$0.141_{\pm 0.778}$	$0.053_{\pm 0.629}$	$0.374_{\pm 0.585}$	357.00MB
DAB-DETR [20]	$0.284_{\pm 0.424}$	$0.159_{\pm 0.395}$	$\mathit{0.650}_{\pm \mathit{0.402}}$	366.90MB
Ours	$\mathbf{0.624}_{\pm \mathbf{0.015}}$	$\mathbf{0.303}_{\pm \mathbf{0.012}}$	$0.632_{\pm 0.026}$	**3.97MB**
Ours(PlantVillage)	$0.984_{\pm 0.193}$	$0.971_{\pm 0.041}$	$0.955_{\pm 0.253}$	3.95MB

space requirement as high as 520 MB, a size approximately 130 times that of the model developed in this study. Although the proposed model slightly lags behind YOLOv5 and DAB-DETR in recall rate with a score of 0.632, it demonstrates significant advantages in model lightweighting. Compared to the YOLO-World and SSD models, the weight file size of this model is reduced by 129.98% and 22.17%, respectively. The substantial optimization in model size enhances its suitability for deployment on edge devices characterized by limited computational capabilities. This development significantly aligns the model with the practical requirements and operational demands of the agricultural sector. Furthermore, the HSFPN-Det model's test results on the PlantVillage dataset are delineated in the concluding row of Table 2. The model demonstrates excellent performance across multiple metrics: mAP@0.5 is 0.984, mAP@0.5:0.95 is 0.971, and the recall rate reaches as high as 0.955. The corresponding standard deviations for these metrics are 0.193, 0.041, and 0.253, respectively. With its streamlined model size of merely 3.95 MB, HSFPN-Det exhibits significant potential for deployment within resource-limited settings.

To thoroughly assess the efficacy of the proposed HSFPN-Det model in identifying rice pests and diseases, we systematically calculate a range of key performance metrics. These metrics encompass the $F1$ score, false positive rates, false negative rates, and confusion matrices, each specifically adapted to individual pest and disease categories. This comprehensive evaluation framework offers an in-depth insight into the model's performance across various classification tasks.

As shown in Table 3, the tungro category achieved the highest performance, as demonstrated by an $F1$ score of 0.834, whereas the leaf spot disease category exhibited the lowest performance, with an $F1$ score of 0.361. Furthermore, the HSFPN-Det model demonstrates notable efficiency in real-time object detection tasks, attaining an inference speed of 108 frames per second. Figure 6 displays a confusion matrix with a perfect diagonal pattern, where all off-diagonal

Table 3 Evaluation indicators for four types of pests and diseases and HSFPN-Det

Categories	Precision	Recall	$F1$ score	Specificity	Sensitivity
Leaf spot disease	0.532	0.273	0.361	0.468	0.726
Brownspot	0.701	0.290	0.410	0.290	0.709
Tungro	0.802	0.868	0.834	0.198	0.131
Bacterial blight	0.695	0.720	0.708	0.305	0.279
HSFPN-Det	0.702	0.632	0.597	0.298	0.481

elements are zero, indicating the absence of inter-class confusion or misclassification. Across a total of 361 samples, the model attained an impeccable accuracy of 100%. This outstanding classification performance underscores the model's exceptional reliability and practical utility in the domain of plant disease recognition. It evidences the model capability for accurate identification of various plant disease categories, thereby offering robust technical support for agricultural disease diagnostics.

4.4 Ablation studies

The ablation experiments conducted in this research unequivocally highlight the exceptional detection accuracy of the constructed model, while concurrently establishing the supremacy of the enhanced neck network architecture over alternative neck network designs. The precise experimental data underpinning these assertions are comprehensively detailed in the statistical results found in Table 4. Through meticulous ablation comparison experiments involving various methodologies, including VanillaNet+BiFPN, HATHead, SETNetV2, DySample, RepHead, AFPN, and CGNet, this study robustly affirms the substantial efficacy of the novel enhancements to the neck network structure for the challenge of rice pest and disease recognition. In parallel, the experimental design adeptly delineates the improvement strategies applied across three distinct modules: the Backbone, Neck, and Head. As indicated by the findings in Table 4, within the context of the dataset employed herein, the refinement and enhancement of the neck network are capable of simultaneously elevating the critical metrics of mAP@0.5 and mAP@.5:.95, while also providing modest augmentation to the model's recall performance.

The VanillaNet architecture combined with BiFPN is 0.590 on the mAP@0.5 metric and is 0.283 on the mAP @0.5:0.95 metric. These results indicate performance deficits of 5.76% and 7.06% when compared to the model introduced in the present study. VanillaNet embodies a simplified design ethos, prioritizing reduced model complexity by excising redundant network depth, skip connections, and self-attention mechanisms. This design philosophy renders it particularly adept for deployment in scenarios constrained by limited computational resources. Nevertheless, owing to its pared-down network architecture and comparatively fewer network layers, the model's capacity for nonlinear expression is restrained. This limitation adversely impacts overall performance, precluding the achievement of ideal detection outcomes on the dataset utilized in this study. Conversely, the SENetV2 model manifests remarkable recall rate performance, achieving a superior value of 0.654, which exceeds that of the model introduced in this study. SENetV2 enhances the global feature representation learning capability through a strategic design involving multi-branch fully connected layers. The core strength of SENetV2 resides in its ability to considerably boost overall performance while incurring only a marginal increase in parameter count.

Table 5 delineates the results of a comparative analysis of various attention mechanisms applied to our dataset, encompassing ACmix [28], MSDA [29], Hybrid Attention Transformer [13], TripletAttention [30], and EMAttention [31]. Although these attention mechanisms yield incremental improvements in overall performance, considerations imperative to deployment–such as parameter count and computational complexity–play a pivotal role in determining model feasibility within environments constrained by limited resources. Our innovative attention mechanism sets itself apart by surpassing existing methodologies in a range of evaluation metrics while also exhibiting remarkable suitability for implementation on devices with limited computational resources.

Based on the preceding analysis, we confidently conclude that our proposed model outperforms other algorithms in detecting rice pests and diseases. More importantly, its compact size makes it more suitable for real-world agricultural environments, enabling deployment on plant protection machinery.

The increase in model parameters and storage requirements resulting from the incorporation of various attention mechanisms is generally minimal; however, their contribution to improving detection accuracy is considerable. This research successfully integrates the Hybrid Attention Transformer mechanism within the HSFPN-Det framework, yielding superior detection accuracy compared to all other evaluated attention mechanisms. The notable effectiveness of the Hybrid Attention Transformer can be primarily attributed to its distinctive three-stage architectural design, which

Table 4 Ablation experiments on the rice pest and disease dataset. [Key: **Best**, *Second Best*]

Methods	Backbone	Neck	Head	mAP@0.5	mAP@.5:.95	Recall
YOLOv8				0.605	0.291	0.623
+VanillaNet + BiFPN [21]	✓	✓		0.590	0.283	0.622
+HATHead [22]			✓	0.611	0.292	*0.642*
+SENetV2 [23]	✓			*0.619*	*0.299*	**0.654**
+DySample [24]				0.609	0.294	0.635
+RepHead [25]			✓	0.617	*0.299*	0.633
+AFPN [26]			✓	0.616	0.291	0.627
+CGNet [27]	✓			0.614	0.292	0.637
+Ours		✓		**0.624**	**0.303**	0.632

Table 5 Comparative results of performance across various attention mechanisms. [Key: **Best**, *Second Best*]

Name	mAP@0.5	mAP@.5:.95	Recall	Precision	GFLOPs
ACmix	*0.624*	**0.304**	*0.622*	0.612	**8.7**
MSDA	0.623	0.299	*0.622*	0.612	*8.3*
TripletAttention	0.625	*0.303*	0.621	*0.623*	8.1
EMAttention	0.619	0.302	0.621	0.618	*8.3*
Hybrid attention transformer	**0.628**	*0.303*	**0.640**	**0.630**	8.1

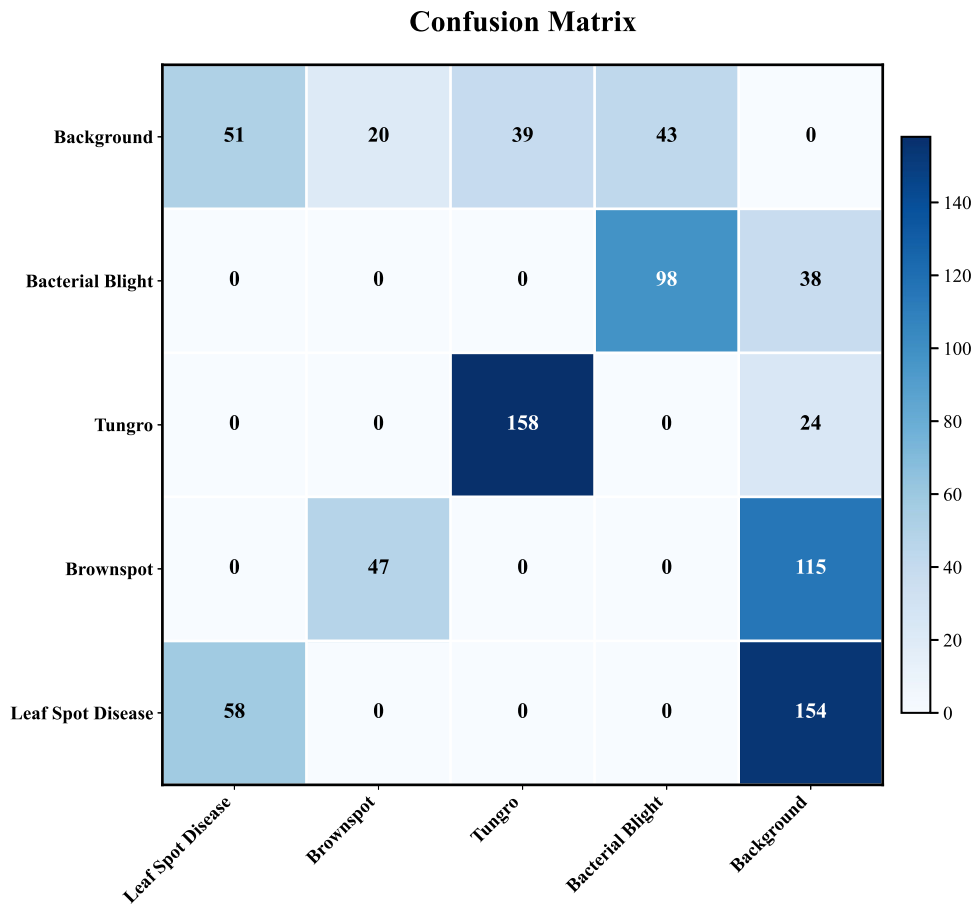

Fig. 6 Confusion matrix for four types of pests and diseases

includes shallow feature extraction, deep feature extraction, and image reconstruction phases. A key component of this architecture is the deep feature extraction module, which consists of multiple Residual Hybrid Attention Groups (RHAGs). Each RHAG integrates several hybrid attention blocks alongside an Overlapping Cross-Attention Block (OCAB). Within each hierarchical attention block, the implementation of channel attention blocks and windowed multi-head self-attention mechanisms enables the effective capture of inter-channel dependencies and spatial positional relationships during feature extraction. The OCAB further enhances feature interaction across window regions, thereby facilitating a more comprehensive and nuanced understanding of contextual information. The image reconstruction stage is critical for transforming the processed features from multiple RHAGs into high-resolution representations, ensuring that these refined feature maps support accurate object localization and classification in subsequent tasks. Experimental results, as presented in Table 5, demonstrate that the integration of the Hybrid Attention Transformer substantially improves the accuracy of rice pest and disease recognition, exhibiting strong performance on the benchmark dataset utilized in this study.

Moreover, as demonstrated in Table 5, the performance of EMAttention with respect to mAP@0.5 is inferior to that of ACmix, MSDA, and TripletAttention, achieving a modest mAP@0.5 value of 0.619. This result is substantially lower than the scores obtained by the other attention mechanisms. However, since the input data in this study are images, the increased computational overhead associated with processing along the channel dimension diminishes the effectiveness of EMA, resulting in suboptimal performance in the experimental assessments.

5 Failure cases

Although the HSFPN-Det proposed in this paper outperforms other methods in comparative experiments, there are still some failure cases in practical applications. The observed performance discrepancy may originate from the model's effectiveness being predominantly validated under controlled experimental conditions, which may not seamlessly translate to the complexities of real-world agricultural environments. In these natural settings, the variability and unpredictability inherent in pest and disease detection pose significant challenges that controlled conditions may fail to adequately simulate or address. Consequently, while a model may demonstrate impressive results in a laboratory setting, its applicability and reliability in practical, dynamic agricultural scenarios require further evaluation and adaptation. Factors such as varying lighting conditions, shadow interference, and adverse weather (e.g., rain, fog, high brightness and occlusion) significantly degrade detection accuracy. This environmental variability contributes to the model's reduced performance when transitioning from laboratory to field applications. Although the HSFPN-Det model is compact-occupying only 3.97 MB-its inference speed and real-time capabilities on edge devices require further optimization. This is particularly critical for deployment on drones or handheld devices used across large-scale farmland, where computational constraints and limited battery life remain substantial bottlenecks.

Furthermore, the model's detection accuracy diminishes when it encounters multiple co-occurring plant diseases, as it struggles to differentiate overlapping symptomatic regions effectively.

To approximate real-world rice growth environments, data augmentation techniques were utilized to simulate four distinct environmental conditions: occlusion, fog, rain, and high brightness, as depicted in Fig. 7. Subsequently, the detection performance of the HSFPN-Det model was assessed under each of these conditions through a series of controlled experimental evaluations.

The outcomes of the comparative analyses are illustrated in Fig. 8. The confidence levels for the original images are 0.58, 0.66, 0.30, and 0.87, which are significantly higher than the values obtained under four different environmental conditions. Notably, no pests or diseases were detected in high-brightness environments, and several instances of missed detections occurred under foggy and rainy conditions. These findings suggest that the detection accuracy of HSFPN-Det decreases when exposed to diverse environmental scenarios. This performance degradation may be attributed to the model being primarily trained on annotated images captured in controlled greenhouse environments, which lack diverse data noise and environmental variability. Future research will focus on enhancing the robustness of the model for pest and disease detection in real-world field environments.

5.1 Real-world image detection

From Figs. 9 and 10, several conclusions can be drawn: YOLOv5 failed to detect specific pests and diseases in the first image, and SSD did not detect any in the first and second images. EfficientDet-D1 detected one pest and disease with a confidence level of 0.69. This paper's proposed model displayed two detection boxes with confidence levels of 0.63 and 0.32, which were marginally lower than those of EfficientDet-D1.

In the identification of the second pest and disease image, both YOLOv8 and the model proposed in this study demonstrated a detection confidence level of 0.94, whereas YOLOv11 exhibited slightly lower confidence levels of 0.94 and 0.91, respectively. In contrast, EfficientDet-D1 recorded

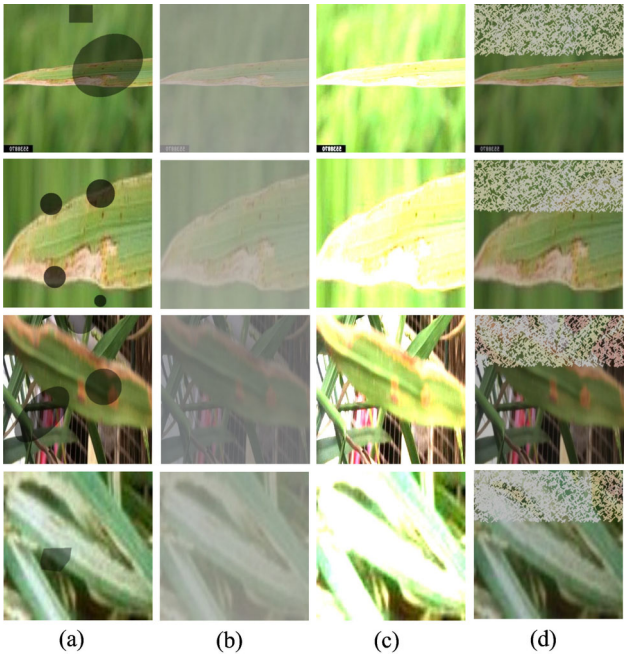

Fig. 7 Images were captured under four distinct environmental conditions: **a** occlusion, **b** fog, **c** high brightness and **d** rain

the lowest confidence level of 0.78 among all evaluated models. Similarly, in the evaluation of the final image, SSD maintained a superior confidence level of 0.89, significantly exceeding the scores of 0.70 and 0.30 achieved by the model introduced herein. When compared to the proposed HSFPN-Det approach, the real-world detection performances of TOOD, DINO, Deformable DETR, and DAB-DETR were substantially lower. Nevertheless, considering the overall detection outcomes and confidence levels, the model presented in this study outperformed both SSD and YOLOv5 in pest and disease detection tasks. Furthermore, the compact architecture and adaptability of the HSFPN-Det model enhance its suitability for deployment in practical agricultural settings.

6 Discussion

6.1 Application of HSFPN-Det in rice pests and diseases detection

The experimental findings demonstrate that the HSFPN-Det model attains outstanding performance across key evaluation metrics, achieving a mean Average Precision (mAP) at IoU threshold 0.5 of 0.624 and an mAP averaged over IoU thresholds from 0.5 to 0.95 of 0.303. This corresponds to a substantial compression rate of 49.87% compared to analogous models. Such a novel advancement presents considerable potential for application in modern agricultural practices. In the context of early and accurate disease detection, the model exhibits a remarkable capability for the rapid identification of major rice diseases, including leaf blight, brown spot, rice blast, and bacterial wilt. By providing timely warning information, it promises significant improvements in disease management and mitigation strategies within rice cultivation. This early detection capacity directly informs the optimal timing for pesticide application, thereby preventing excessive use of agrochemicals. Consequently, this not only reduces production costs but also substantially diminishes adverse environmental impacts, aligning well with the principles of sustainable agricultural development. Regarding practical deployment, the model's lightweight architecture enables efficient operation on edge computing devices with limited computational resources, such as portable detection instruments or agricultural drone platforms, facilitating real-time monitoring during field operations. This technological innovation effectively addresses challenges associated with the scarcity of specialized technical personnel and low detection efficiency in conventional agricultural settings, thereby enhancing the timeliness and scope of pest and disease control measures. In terms of detection accuracy, it slightly surpasses the recently released YOLOv9 and YOLOv8 models in mAP@0.5, and achieves a 3.1% improvement in mAP@0.5:0.95 relative to YOLOv9. This superior accuracy is primarily attributed to its novel High-Selective Feature Pyramid Network architecture, which more effectively integrates and exploits multi-scale image feature information. With respect to computational resource efficiency, HSFPN-Det's performance is particularly notable.

Compared to YOLO-World and SSD, it requires only 3.97 MB of storage, corresponding to compression rates of 99.2% and 95.7%, respectively, while maintaining or even enhancing detection accuracy. This extreme lightweight design renders HSFPN-Det an optimal solution for deployment scenarios constrained by limited computational capacity.

In real-world testing, HSFPN-Det outperforms comparative models in robustness under complex backgrounds and lighting conditions, which is crucial for field applications. The model can effectively identify rice diseases in complex natural environments, maintaining a high level of confidence (e.g., achieving dual-target detection of 0.63 and 0.32 in tests). As rice is a major global food crop, pests and diseases cause an annual yield reduction rate of 10%–15%. HSFPN-Det's remarkable accuracy and real-time operational capacity have the potential to mitigate crop loss by an estimated 40%–60%. As evidenced by experimental findings, the model consistently delivers robust performance in real-world environments, most notably excelling in the precise identification of brown spot and bacterial wilt. This heightened level of accuracy empowers farmers to implement timely intervention strategies, thereby optimizing disease management and significantly enhancing agricultural productivity. Addition-

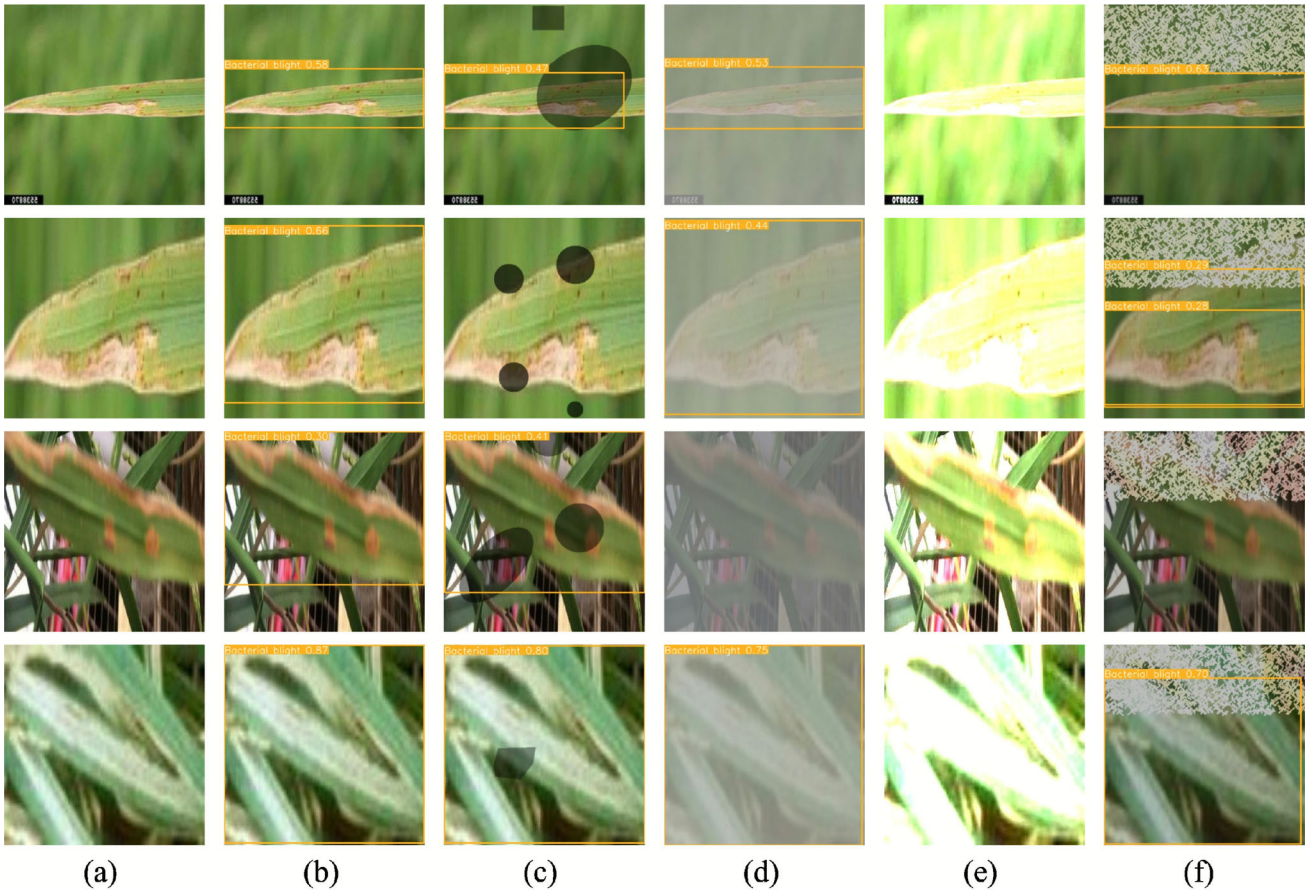

Fig. 8 The visualization encompasses comparative experiments performed across four discrete environmental conditions–namely occlusion, fog, rain, and heightened brightness. The figure is organized as follows: **a** portrays the original image; **b** illustrates the detection outcome on the original image; **c** reveals the detection result when confronted with occlusion; **d** presents the detection result within foggy conditions; **e** exhibits the detection result under conditions of elevated brightness; and **f** demonstrates the detection result amidst rainy conditions. This structured presentation allows for an insightful comparison of the model's adaptability and performance across a range of challenging environmental scenarios

ally, the model's lightweight design (3.97MB) significantly reduces deployment costs, making it suitable for small-scale farmers in developing countries. Calculations show that compared to traditional manual identification methods, using HSFPN-Det can shorten pest and disease detection time by 85% while increasing identification accuracy by more than 30%, directly impacting the optimal timing for prevention and improving crop yield and quality.

6.2 HSFPN-Det's performance in other crops or pests beyond rice

The principal advantage of the HSFPN-Det model resides in its seamless integration of the High-level Feature Selection Pyramid Network, a deformable self-attention module, and the Hybrid Attention Transformer. Together, these sophisticated components substantially augment the model's proficiency in discerning target features amidst a multitude of complex backgrounds. This amalgamation of advanced techniques confers a robust ability to navigate and interpret intricate environmental variations, thereby enhancing overall detection accuracy and reliability. This model is anticipated to perform well in detecting major crops such as wheat, corn, and soybeans. For wheat, common ailments such as fusarium head blight, rust, and powdery mildew exhibit visual parallels to rice diseases, manifesting as spots or discolored areas on the leaves. The multi-scale feature fusion capability of HSFPN-Det can effectively capture these attributes. Of note, the orange-red raised lesions associated with wheat rust have distinct color characteristics, rendering the model's Hybrid Attention transformer potentially highly sensitive to these features. In detecting corn pests and diseases, the identification of northern leaf blight, southern leaf blight, and corn borers could be enhanced by the HSFPN-Det model's design. Corn diseases typically present as extensive areas of lesions or stripes on the leaves, and the HSFPN module's effective

Fig. 9 Rice pest and disease detection results under real-world conditions, from top to bottom, are Original images, YOLOv8, YOLOv5, YOLO-World, YOLOv11 and SSD

fusion capability for different scale features assists in precisely identifying these symptoms.

6.3 Limitation

While the HSFPN-Det model demonstrates superior performance, it still faces several key challenges in practical application environments. Firstly, the impact of environmental factors on detection effectiveness is a significant issue. The dataset in the paper is primarily collected in controlled experimental environments; however, variations in lighting, shadow effects, and weather conditions (such as rain and fog) in actual farmland environments can significantly reduce detection performance. Such environmental variability poses the risk of diminishing the model's performance as it transitions from the controlled conditions of the laboratory to the

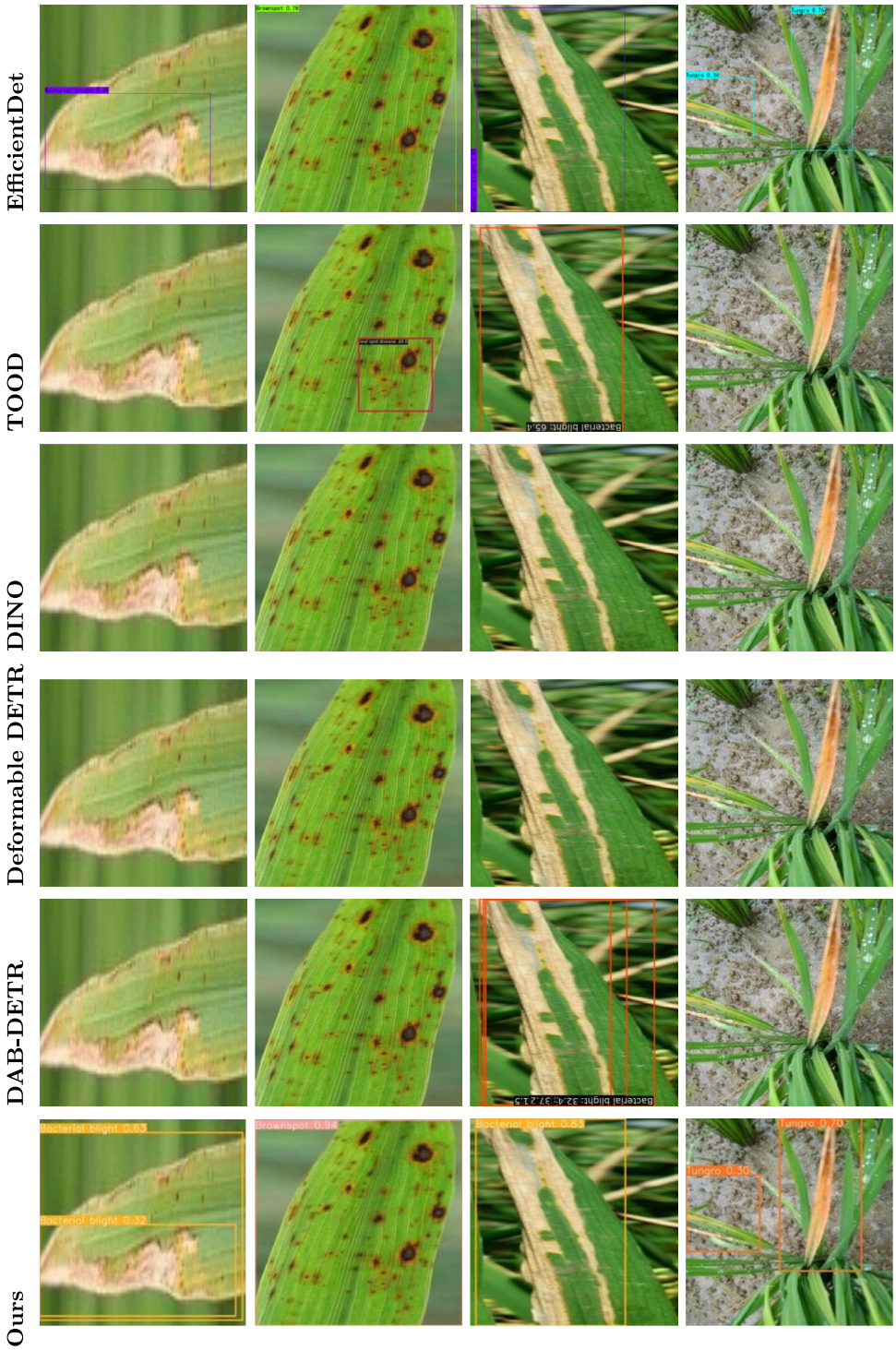

Fig. 10 Rice pest and disease detection results under real-world conditions, from top to bottom, are EfficientDet, TOOD, DINO, Deformable DETR, DAB-DETR and the proposed method

dynamic and unpredictable settings of real agricultural environments. This shift can introduce an array of challenges that potentially affect the model's efficacy, necessitating adaptive strategies to maintain its operational robustness and accuracy in practical applications. Another challenge is the recognition of different stages of disease development. The same disease at different stages may exhibit different visual characteristics, and existing models may have limited capabilities in early disease detection. Insufficient sensitivity to mild symptoms may delay farmers' intervention timing, affecting yield.

Additionally, the model's ability to handle overlapping disease areas is limited, and detection accuracy can be affected when multiple diseases appear on a single plant simultaneously. Technically, although the HSFPN-Det model

is smaller in size compared to other models (only 3.97MB), the balance between inference speed and real-time performance still needs optimization. For edge devices, especially drones or handheld devices used in large farmland applications, the limitations of computational resources and battery capacity remain a bottleneck for practical deployment.

To address the challenges outlined above, this research delineates several strategic improvements aimed at enhancing the model's performance. These proposed refinements are intended to bolster the model's robustness and adaptability, ensuring sustained efficacy in diverse and unpredictable agricultural environments. (a) Dataset expansion and environmental adaptability enhancement. The training dataset can be enriched by collecting image samples under more diverse environmental conditions, which should cover scenarios with various lighting intensities, weather conditions, and background complexities. Meanwhile, domain adaptation techniques are employed to help the model better transfer from laboratory-controlled conditions to real-world farmland environments. (b) Further improvement of attention mechanisms. The existing Hybrid Attention Transformer architecture is deeply optimized to enhance its ability to distinguish overlapping disease regions while increasing the model's sensitivity to capturing minute disease features. (c) Application of model lightweighting and inference acceleration techniques. By implementing advanced technologies such as knowledge distillation, neural network pruning, and model quantification, the model size is further reduced and inference speed is improved without compromising detection accuracy. Through the implementation of these technical optimization measures, the applicability and practical value of the HSFPN-Det model in real agricultural production environments will be significantly enhanced, enabling it to become a truly reliable smart agriculture solution that contributes to advancing precision agriculture development and achieving sustainable crop protection goals.

7 Conclusion

In this paper, we propose HSFPN-Det, an efficient model for rice pest and disease detection. Through the integration of multi-scale feature fusion methodologies alongside deformable self-attention mechanisms, we have achieved notable advancements in the model's image feature extraction prowess. This refined model not only embodies a more streamlined design but also demonstrates significant improvements in detection accuracy while concurrently reducing the demand on computational resources. These attributes render the model exceptionally apt for deployment on plant protection equipment constrained by computing power limitations. An exhaustive evaluation was conducted, scrutinizing the performance of comparison models across two pivotal metrics: mAP@0.5 and mAP@.5:.95. The model's capability for real-time pest and disease detection serves as a potent technical instrument for agricultural experts, thereby advancing the frontier of precision crop health monitoring. This technological innovation facilitates the implementation of prompt phytosanitary interventions and is particularly valuable for integrated pest management systems that necessitate swift action. The ramifications of these advancements are profound for modern agriculture. On one hand, the HSFPN-Det model excels in the early-stage identification of prevalent rice diseases, such as leaf spot and brown spot, thus providing a scientific foundation for timely pesticide application. This precocious identification ability significantly curtails pesticide overuse, diminishing agricultural production expenses and mitigating adverse environmental impacts. On the other hand, the model's streamlined design guarantees stable operation on edge computing devices characterized by constrained computational resources, including portable detection devices and agricultural drone systems. This capability for real-time field detection markedly heightens the responsiveness and operational breadth of pest and disease management operations, effectively addressing the hindrances posed by scarce agricultural expertise and the inefficiencies inherent in conventional detection approaches.

Acknowledgements This work was supported by the Open Research Fund of Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion, Wuhan Textile University (Grant Number: 2023HBITF02), the Open Research Fund of Key Laboratory of Ethnic Language Intelligent Analysis and Security Management of MOE, Minzu University of China (Grant Number: ORP-202406).

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by YangYang and Jianlin Zhu. The first draft of the manuscript was written by YangYang, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Code and datasets are available at https://github.com/Louaq/HSFPN-Det

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Watt, M.S., Holdaway, A., Watt, P., Pearse, G.D., Palmer, M.E., Steer, B.S., Camarretta, N., McLay, E., Fraser, S.: Early prediction of regional red needle cast outbreaks using climatic data trends and satellite-derived observations. Remote Sens. **16**(8), 1401 (2024)
2. Xia, Y., Che, T., Meng, J., Hu, J., Qiao, G., Liu, W., Kang, J., Tang, W.: Detection of surface defects for maize seeds based on yolov5. J. Stored Prod. Res. **105**, 102242 (2024)

3. Cai, X., Chen, Z., Sheng, B.: Spt: swin pyramid transformer for object detection of remote sensing. Compute. Sci. **50**(1), 105–113 (2023)
4. Bingyan, Z., Zhihua, C., Sheng, B.: Remote sensing image detection based on perceptually enhanced swin transformer. Comput. Eng. **50**(1), 216–223 (2024)
5. Chen, Y., Zhang, C., Chen, B., Huang, Y., Sun, Y., Wang, C., Fu, X., Dai, Y., Qin, F., Peng, Y.: Accurate leukocyte detection based on deformable-detr and multi-level feature fusion for aiding diagnosis of blood diseases. Comput. Biol. Med. **170**, 107917 (2024)
6. Guo, L., Wu, Y., Zhao, J., Yang, Z., Tian, Z., Yin, Y., Dong, S.: Rice disease detection based on improved yolov8n. In: 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL), pp. 123–132. IEEE (2025)
7. Peng, H., Yao, L., Liu, H., Peng, S., He, H., Xu, H., Li, M.: Different life cycles of rice pests' images recognition based on adaptive lightweight dc-ghost module. Expert Syst. Appl. **255**, 124645 (2024)
8. Zheng, Y., Zheng, W., Du, X.: Paddy-yolo: an accurate method for rice pest detection. Comput. Electron. Agric. **238**, 110777 (2025)
9. Bai-yi, S., Jiong, M., Hao-yang, Y., Hong-jie, W., jie, Y.: Detection of ears of rice in field based on ssd. In: Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, pp. 228–232 (2020)
10. Li, Z., Shen, Y., Tang, J., Zhao, J., Chen, Q., Zou, H., Kuang, Y.: Imll-detr: an intelligent model for detecting multi-scale litchi leaf diseases and pests in complex agricultural environments. Expert Syst. Appl. **273**, 126816 (2025)
11. Shahbaz, A., Jo, K.-H.: Deep atrous spatial features-based supervised foreground detection algorithm for industrial surveillance systems. IEEE Trans. Industr. Inf. **17**(7), 4818–4826 (2020)
12. Nguyen, D.-K., Ju, J., Booij, O., Oswald, M.R., Snoek, C.G.: Boxer: Box-attention for 2d and 3d transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4773–4782 (2022)
13. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22367–22377 (2023)
14. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16901–16911 (2024)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37. Springer (2016)
16. Tan, M., Pang, R., Le, Q.V.: Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
17. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: Tood: task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3490–3499. IEEE Computer Society (2021)
18. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y.: Dino: detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
19. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
20. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (2022)
21. Chen, H., Wang, Y., Guo, J., Tao, D.: Vanillanet: the power of minimalism in deep learning. ArXiv:2305.12972 (2023)
22. Yu, H., Wan, C., Liu, M., Chen, D., Xiao, B., Dai, X.: Real-time image segmentation via hybrid convolutional-transformer architecture search. ArXiv:2403.10413 (2024)
23. Narayanan, M.: Senetv2: aggregated dense layer for channelwise and global representations. ArXiv:2311.10807 (2023)
24. Liu, W., Lu, H., Fu, H., Cao, Z.: Learning to upsample by learning to sample. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6027–6037 (2023)
25. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)
26. Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., Liang, R.: Afpn: asymptotic feature pyramid network for object detection. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2184–2189. IEEE (2023)
27. Wu, T., Tang, S., Zhang, R., Cao, J., Zhang, Y.: Cgnet: a light-weight context guided network for semantic segmentation. IEEE Trans. Image Process. **30**, 1169–1179 (2020)
28. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G.: On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 815–825 (2022)
29. Jiao, J., Tang, Y.-M., Lin, K.-Y., Gao, Y., Ma, A.J., Wang, Y., Zheng, W.-S.: Dilateformer: multi-scale dilated transformer for visual recognition. IEEE Trans. Multimedia **25**, 8906–8919 (2023)
30. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3139–3148 (2021)
31. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z.: Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)

Yang Yang received the B.S. degree in computer science from the South-Central Minzu University, Wuhan, China, in 2024. He is currently pursuing the master's degree with South-Central Minzu University. His research interests focus on medical image segmentation and computer vision.

Yuxin Hong received the B.Eng. degree in data science and big data technology from Huazhong Agricultural University, Wuhan, China, in 2022, and the M.Eng. degree in computer science and technology from South-Central Minzu University, Wuhan, China, in 2025. From 2023 to 2024, she was a research intern at the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. She is pursuing her Ph.D. degree at Lingnan University, Hong Kong. Her research interests include computer vision and data-centric efficiency.

Wenjie Yu ,an undergraduate of the 2022 cohort, is majoring in software engineering at South-Central Minzu University. His research interests include image recognition and segmentation.

Xiao Zhang received the B.Eng. and M.Eng. degrees from the South-Central Minzu University, Wuhan, China, in 2009 and 2011, respectively, and the Ph.D. degree from Department of Computer Science in City University of Hong Kong, Hong Kong, 2016. He was a visiting scholar with Utah State University, Utah, USA, and University of Lethbridge, Alberta, Canada. During 2016–2019, he was a Postdoc Research Fellow at Singapore University of Technology and Design. Currently, he is associate professor with the College of Computer Science, South-Central Minzu University, China. His research interests include wireless and UAV networking, algorithms design and analysis, and combinatorial optimization.

Bo Yang received the B.S. degree in computer science and technology from the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), in 2001, the M.S. degree in water conservancy and hydroelectric engineering, and the Ph.D. degree in spatial information science and technology from HUST, in 2004 and 2008, respectively. From 2008 to 2011, she worked as a postdoctoral research fellow at HUST. Since 2011, she has been affiliated with South-Central Minzu University (SCMU), where her research interests span computer modeling and simulation, GIS, computer vision, and sign language recognition.

Meng Shi received the B.E. degree in thermal energy and power engineering from the Huazhong University of Science and Technology in 2016 and Ph.D. degree in architecture and civil engineering from City University of Hong Kong. She is currently an Associate Professor with School of Computer Science, South-Central Minzu University. Her research interests include emergency evacuation dynamics and intelligent transportation.

Yangguang Sun received the M.Sc. degree in computational mathematics and the Ph.D. degree in pattern recognition intelligence system, from the Huazhong University of Science and Technology, China, in 2005 and 2009, respectively. He is currently a vice professor in the School of Computer Science, South-Central Minzu University, China. His current research interests are image processing and computer vision.

Jun Wang received the Ph.D. degree in computer science from Wuhan University, China, in 2017. He is currently a lecturer with School of Computer Science (School of Artificial Intelligence), South-Central Minzu University, Wuhan, China. His research interests include network security, privacy protection, and AI security in AIoT.

Jianlin Zhu received her Ph.D. degree in radio physics from the Huazhong Normal University, Wuhan, China, in 2013. She is currently a Lecturer and Master's supervisor with the College of Computer Science, South-Central Minzu University, Wuhan, China. Her research interests include computer vision, virtual simulation, and smart health care.