Computers and Electronics in Agriculture 227 (2024) 109552

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

High-throughput spike detection and refined segmentation for wheat Fusarium Head Blight in complex field environments

Qiong Zhou [a,b,c], Ziliang Huang [a,b], Liu Liu [d,*], Fenmei Wang [a,b], Yue Teng [e], Haiyun Liu [a,b], Youhua Zhang [c], Rujing Wang [a,b,*]

[a] Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
[b] University of Science and Technology of China, Hefei 230026, China
[c] School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China
[d] Hefei University of Technology, School of Computer Science and Information Engineering, Hefei 230009, China
[e] Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei 230071, China

ARTICLE INFO

Keywords:
High-throughput detection
Wheat spike
Fusarium head blight
Instance segmentation
Fine-grained understanding

ABSTRACT

Fusarium Head Blight (FHB) is a devastating disease of wheat worldwide. It is an explosive epidemic disease that can severely reduce or even fail wheat production. Estimating the disease ear rate and disease severity is crucial for effective plant protection. Manual assessment is labor-intensive and time-consuming. Accurately and quickly segmenting wheat ears and areas affected by Fusarium head blight (FHB) in complex field environments is essential for quantitative assessment of wheat trait phenotypes and FHB in wheat plants. This paper presents DeepFHB, an automated method for efficiently detecting, locating, and segmenting dense wheat spikes and diseased areas in digital images captured under natural field conditions. The experiment consists of three steps: Firstly, the process begins by generating initial coarse-grained mask predictions at lower resolutions to provide a rough segmentation. Secondly, a quadtree-based method is employed to identify and refine multi-scale inconsistent regions. Finally, a transformer-based refinement network is introduced to predict highly accurate instance segmentation masks. The results demonstrate that the DeepFHB algorithm outperforms traditional methods in detecting and segmenting diseased areas. Our DeepFHB model achieves state-of-the-art single-model results of 64.408 box AP and 64.966 mask AP on the FHB-SA dataset. This study is capable of rapidly and accurately segmenting wheat spikes and wheat scab lesions in agricultural scenarios with high field density, high crop occlusion, and high background interference. This provides a foundation for subsequent targeted research to assist agricultural workers in assessing the severity of wheat diseases.

1. Introduction

Wheat is one of the most important cereal crops in the world (Zhang et al., 2022a), with over 40 % of the global population relying on it as a primary food source. Ensuring the steady increase in wheat production is essential for global food security. Wheat production is constantly threatened by various diseases, pest infestations, and abiotic stress. Among these, Fusarium head blight, commonly referred to as scab, is a prevalent affliction in wheat caused by fungi such as Fusarium graminearum, which significantly impacts wheat yield (Gao et al., 2022). The disease can reduce wheat production by infecting the wheat spikes during the flowering stage, leading to red or black mold spots during the grain-filling and ripening stages. Fusarium head blight also produces mycotoxins, which can cause food poisoning in humans and animals when consumed. This disease is particularly severe in the Yangtze and Huai River regions of China. The most common method for monitoring and identifying Fusarium head blight (FHB) is manual observation. This traditional assessment method is time-consuming, labor-intensive, and prone to human error (Zhang et al., 2019). Therefore, there is an urgent need for more effective and precise methods to evaluate diseases in wheat cultivation. Fine-grained understanding and analysis for wheat ear disease assessment and accurate yield estimation are necessary. Refining segmentation of wheat spikes and areas affected by FHB is crucial for quantifying wheat traits and assessing the impact of FHB on wheat plants.

Current research on wheat primarily focused on the identification of

* Corresponding authors at: Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China (R. Wang).
E-mail addresses: liuliu@hfut.edu.cn (L. Liu), rjwang@iim.ac.cn (R. Wang).

https://doi.org/10.1016/j.compag.2024.109552
Received 28 January 2024; Received in revised form 12 September 2024; Accepted 11 October 2024
Available online 18 October 2024

wheat ear diseases, wheat head detection, and counting (Gao et al., 2024; Liu et al., 2022; Zhao et al., 2023; Zhou et al., 2022). While hyperspectral imaging has proven highly effective for early detection of diseases and pests, capturing subtle physiological and biochemical changes in plant leaves (Jin et al., 2018), it faces significant limitations. The large volume of hyperspectral data requires complex processing and analysis, posing challenges for real-time application in rapid-response scenarios. The necessity for specific field equipment like calibration plates and the high costs associated further restrict its widespread adoption. With the rapid development of deep learning technology, its applications in various fields (Garg et al., 2021) have shown remarkable results. Deep learning offers a promising non-destructive alternative for detecting wheat heads and diseases. It has emerged as a key tool in agricultural applications, particularly for identifying wheat ear diseases, an area that has seen considerable research interest. For instance, Bao et al. (2021) proposed a lightweight convolutional neural network (CNN) framework for automated detection of wheat ear diseases like glume blotch and scab under natural field conditions. Gu et al. (2021) developed a feature fusion approach that integrates deep convolutional features with shallow features extracted from high-resolution digital RGB images of FHB at various disease stages, tailored for controlled indoor conditions. Despite the advancements, the detection in complex backgrounds remains challenging, typically requiring high-throughput capabilities not afforded by traditional methods. Zhang et al. (2022a) advanced this by integrating the Yolov5 object detection network with advanced techniques to evaluate the damage under field conditions effectively. Subsequently, they introduced a Rotating YOLO Wheat Detection network and a Simple Spatial Attention network, capable of detecting wheat head images with detection boxes of arbitrary orientation (Zhang et al., 2023). These studies (Sun et al., 2022; Yang et al., 2021) though innovative, often rely on bounding box outputs which are insufficient for tasks like precise disease assessment and yield prediction, where detailed pixel-level segmentation is crucial. Recognizing this gap, some researchers (Ma et al., 2020) have applied semantic segmentation algorithms in field environments to segment wheat spikes, proposing a method based on semantic segmentation for pixel-level classification, aiming to achieve accurate segmentation of wheat spikes from canopy images captured in field conditions. However, mutual occlusion between wheat spikes, especially in densely planted scenarios, may prevent individual segmentation of wheat spikes. To accomplish counting tasks and separate clustered wheat spikes in semantic segmentation, it is necessary to combine other machine learning algorithms to distinguish different instances of wheat spikes (Zhang et al., 2021; Zhang et al., 2019). In summary, the afore-mentioned detection and segmentation networks exhibit the following issues: Firstly, detection networks aimed at directly obtaining the number of wheat spikes struggle to accurately segment Fusarium Head Blight (FHB) lesion areas. Secondly, Semantic segmentation network methods face challenges in directly obtaining the count of wheat spikes based on segmentation results.

Understanding these limitations is essential for developing more effective methods for wheat spike segmentation and disease detection. Overcoming these challenges may require integrating other technologies or approaches. Precise pixel-level segmentation is crucial for tasks such as disease assessment and yield prediction in wheat management. Therefore, it is necessary to explore segmentation methods that can accurately determine the wheat spike area under real field conditions. Instance segmentation can accurately identify and separate different object instances within an image enabling precise counting of object instances. Previous researches has utilized instance segmentation for spike segmentation (Batin et al., 2023; Zhang et al., 2022b) and wheat disease segmentation (Gao et al., 2022; Qiu et al., 2019; Su et al., 2020). Zhang et al. (2022b) introduced a novel instance segmentation method, employing a Hybrid Task Cascade model, aimed at resolving the problem of wheat spike detection. Batin et al. (2023) presented an instance segmentation approach, rooted in the Cascade Mask RCNN architecture, complemented by model enhancement and hyperparameter optimization for wheat spike segmentation and counting from field imagery. Accurately identifying wheat spikes against complex backgrounds is essential for obtaining image-derived wheat phenotype data, such as yield estimation and morphological characteristics of the spikes. Simultaneously, rapid and precise segmentation of wheat Fusarium Head Blight lesions is crucial for spike disease assessment. This assists agricultural workers in confirming the severity of wheat diseases and conducting subsequent targeted research. Su et al. (2020) developed a dual Mask-RCNN model for rapid segmentation of wheat spikes and FHB diseased areas. Gao et al. (2022) employed an automated tandem dual BlendMask deep learning framework, designed for segmentation of both the wheat spikes and diseased areas, to enable rapid assessment of disease severity. Both of the above methods use a sequential working mode, where one must wait for the previous task to complete before starting the next one. This affects the overall efficiency of the segmentation. It has been observed through extensive experimentation that existing instance segmentation models, such as Mask R-CNN (He et al., 2017), Cascade Mask RCNN (Cai and Vasconcelos, 2018), HTC (Chen et al., 2019), BlendMask (Chen et al., 2020), Swin Transformer (Liu et al., 2021), YOLO (Wang et al., 2021), and others, are not suitable for performing high-throughput wheat spike segmentation and individual spike disease assessment tasks simultaneously. A fine-grained understanding and analysis of wheat spike Fusarium Head Blight (FHB) is essential for wheat management, including disease assessment and yield prediction. A segmentation method is required to quickly and precisely determine wheat spike areas under real field conditions, while detecting and counting both healthy and diseased wheat spikes.

Deep learning-based methods for detecting FHB still face challenges. Our experimental dataset includes pose-variant, overlapping, densely distributed, and differently scaled targets, which negatively affect feature extraction. To address this issue, we propose a novel approach called DeepFHB, inspired by a Mask Transfiner(Ke et al., 2022) model, for high-throughput spike detection and refined segmentation in the context of Wheat Fusarium Head Blight (FHB). Because of the irregular shapes and sizes of wheat spikes and diseased areas, fixed-shape convolutional kernels often perform suboptimally when dealing with such targets. To improve performance, we introduce the use of deformable convolution. Deformable convolution (Dai et al., 2017) is a technique that adds an offset at each sampling point of the standard convolution. The use of deformable convolution allows the convolutional kernel to sample from nearby regions, expanding its receptive field. This effectively mitigates the issue of misalignment of contextual features in segmentation tasks, resulting in enhanced segmentation accuracy. By combining the incoherent region detector and refinement transformer into a single network, refined segmentation can be achieved through an end-to-end method. The major contributions of this article are as follows: (1) A high-throughput deep learning architecture is presented. It can simultaneously detect wheat spikes and disease spots, segment wheat spikes from complex field environments, and isolate Fusarium Head Blight (FHB) spots from the wheat spikes. (2) A coarse-to-fine strategy is employed by our approach, beginning with a multi-scale deep feature pyramid and object detection heads that propose bounding boxes and generate initial coarse masks for segmentation. Inconsistent regions are improved using a quadtree-based method and a lightweight detector. A transformer-based network then enhances the accuracy of instance segmentation. This method offers clear advantages in tasks that require finer object recognition, counting, contour information, and handling occlusion. (3) An end-to-end instance segmentation model is provided by our method, enabling the concurrent execution of multiple tasks, including the detection and segmentation of wheat spikes and diseased areas. Concurrent processing significantly enhances work efficiency while maintaining an average testing time of three seconds per image. This is especially beneficial for managing high-throughput datasets or performing computationally intensive tasks. The paper is structured as follows: Section 2 provides an overview of the

Table 1
The parameters of camera setting.

Variable	Value/State
Camera model	NIKON D5300; HUAWEI ELE-AL00, Xiaomi MI 9
Image size	4496 × 3000;2992 × 2000;3648 × 2736;4000 × 3000
Zoom	No zoom
Flash mode	No flash
Aperture Av.	f/5.3; f/1.8
Focal length	90 mm; 52 mm; 6 mm
Macro	Off
ISO	ISO-400; ISO-50
Image type	JPG

wheat spike dataset used in this research and outlines the proposed methodology for efficient, high-throughput spike detection and refined segmentation for FHB. Section 3 details the algorithm experiments conducted, encompassing various tests, the evaluation of outcomes, and a comparative analysis of the results. Finally, Section 4 summarizes our work.

2. Materials and method

2.1. Image dataset

The field experiment was conducted at a field station in Fengtai County, Huainan City, Anhui Province, China (Latitude: 32°53′11.1163″: N, Longitude: 116°32′6.7273″: E). The original images of wheat spikes were sourced from three viewpoints using a digital camera NIKON D5300 and two mobile phones (HUAWEI ELE-AL00, Xiaomi MI 9), with parameters detailed in Table 1. The growth stage of wheat during image capture is a crucial determinant for effective FHB detection. Identification of affected spikes was not feasible during the early flowering and late maturing stages. The optimal period for disease evaluation was determined to be when spike symptoms are visible, but prior to senescence. The model is utilized during the mid to late stages of the reproductive period of wheat, which includes the Full Flowering, Milk, and Dough stages, extending up to just before the onset of senescence. We anticipate that users will capture images of wheat trial plots with minimal constraints on image acquisition parameters, such as the angle of shot and distance. Consequently, we collected the current dataset under various conditions, including complex backgrounds, diverse distances, shooting angles, and varying illumination, as depicted in Fig. 1. The statistics on the number of various types of training samples under different imaging situations are shown in Table 2. Our dataset includes both simple and complex backgrounds, with a significantly larger number of samples in the complex background category, reflecting our focus on complex field scenes. Regarding diversity distance, shooting angle, and illumination, the sample counts for each subclass in our dataset are relatively balanced, which facilitates accurate recognition and segmentation of objects at different distances, angles, and lighting conditions by the model. This strategy augments the data diversity and subsequently enhances the adaptability and robustness of the model.

We build a wheat spike and FHB segmentation dataset named FHB-SA. A total of 1251 images of wheat with Fusarium head blight (FHB) were captured between 4:00p.m. to 7:00p.m. on May 12th, 2021, under complex wheat field conditions. The original images captured by the camera had a resolution of 4496 × 3000 and 2992 × 2000, while those captured by the mobile phone had a resolution of 3648 × 2736 and 4000 × 3000. To reduce the computational load during training, the original images were resized to 1496 × 1000pixels. The wheat spikes and FHB disease spots were labeled using Labelme software to generate two categories of mask maps, marked as "1″ and "2" respectively (Fig. 2). We structured our dataset into training and validation subsets with the intention of a 9:1 split. However, to ensure statistical robustness and representativeness across different categories, the actual distribution resulted in 1031 images for training and 220 images for validation. This

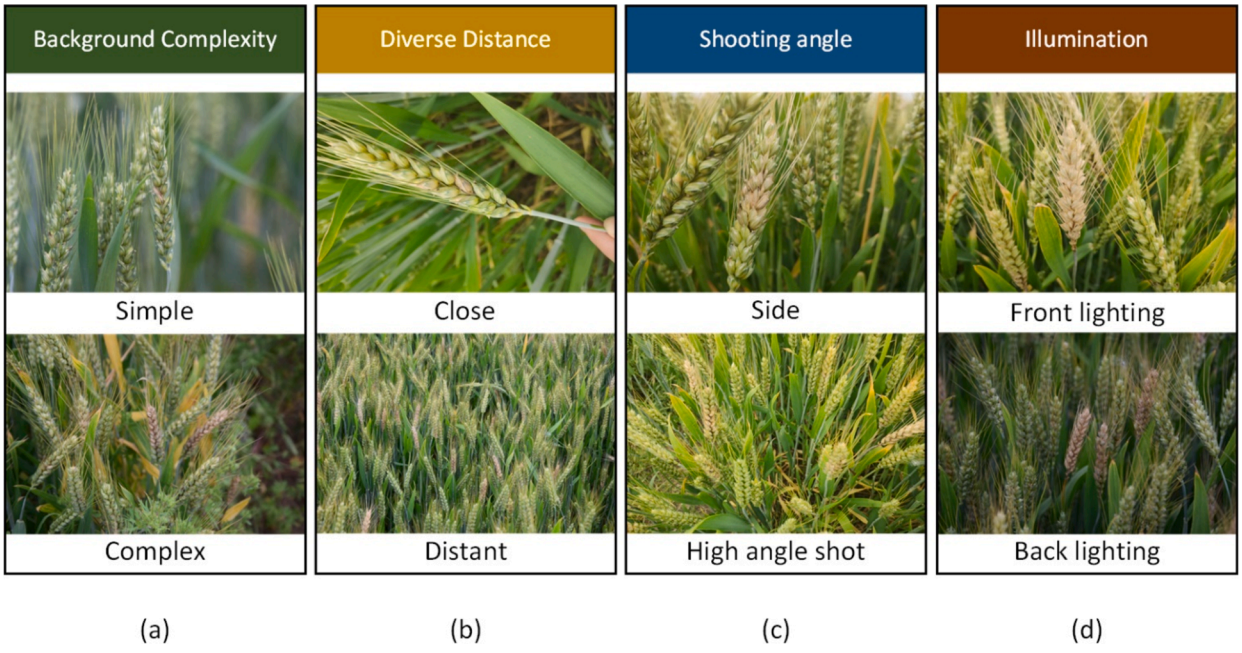

Fig. 1. Data collection under different conditions (a) background complexity (b) diverse distances (c) shooting angel (d) illumination.

Table 2
Statistic on the number of various types of training samples under different imaging situations.

Scene Type	Simple	Complex	Close	Distant	Slide	High angel shot	Front lighting	Back lighting
Number of images	10	1021	494	537	550	481	596	435

Fig. 2. Data annotation by Labelme software (the labeled FHB disease spots are shown as red "1″, and the labeled wheat spikes are shown as green labeled "2").

Table 3
Statistics of Training and Validation Subsets.

Class name	Training			Validation			All		
	Images	Instances	Avg.	Images	Instances	Avg.	Images	Instances	Avg.
Wheat spikes	1031	26124	25.34	220	6270	28.5	1251	32394	25.89
FHB disease spots	1031	7558	7.33	220	1685	7.66	1251	9243	7.39

This table presents the statistics for each class in the training and validation subsets. For each class, the table shows the number of images, the number of instances, and the average number of instances per image.

distribution does not strictly adhere to a 9:1 ratio, reflecting a practical adjustment to meet the specific conditions of our dataset. The training subset, consisting of 1031 images with 26,124 wheat spikes and 7558 Fusarium Head Blight (FHB) disease spots, is used to supervise our model through a rigorous 10-fold cross-validation process for hyper-parameter optimization. The validation subset, used exclusively for testing and not participating in the training process, comprises 220 images with 6270 spikes and 1685 FHB disease spots. This subset functions effectively as a test set, employed to evaluate the system's performance, accuracy, and reliability. The dataset was organized through stratified sampling to ensure representativeness across the different categories of data. Statistical details of our dataset distribution are provided in Table 3.

Fig. 3 provides critical statistical information on the number of instances per image and the proportion of image area occupied by each instance, including wheat spikes and Fusarium Head Blight lesions. A comprehensive count of instances present within each image was performed, and the resulting data is illustrated in Fig. 3(a) and Fig. 3(b). This reveals a notable variability in the quantity of instances per image within our dataset, with the instances ranging from extremely dense (peaking at 200 instances per image) to exceptionally sparse scenarios (with minimal representation of only a single instance per image). Furthermore, it is worth noting that over 75 % of the images within the dataset contain multiple instances, indicating an instance count of three or more. In contrast to traditional disease recognition studies that often focus on single-target identification, our research addresses the more complex challenge of multiple-target detection. As shown in Table 3, the average number of instances per image in our dataset is 25.89, while the average number of FHB disease spots per image is 7.39, highlighting the multi-target nature of our study. Objects within real-field scenarios are predominantly small in dimension. Notably, the size of a staggering 90 % of these targets, encompassing both wheat spikes and disease spots, does not exceed 1 % of the entire image at its maximum extent, as illustrated in Fig. 3(c) and Fig. 3(d). The density and size distribution of instances within the images highlight the real-field challenges our model faces, particularly in accurately identifying and segmenting small and densely clustered objects in complex field scenes. This complexity underscores the need for advanced detection and segmentation methods, which our model aims to address, particularly tailored to handle the intricacies of multi-target scenarios.

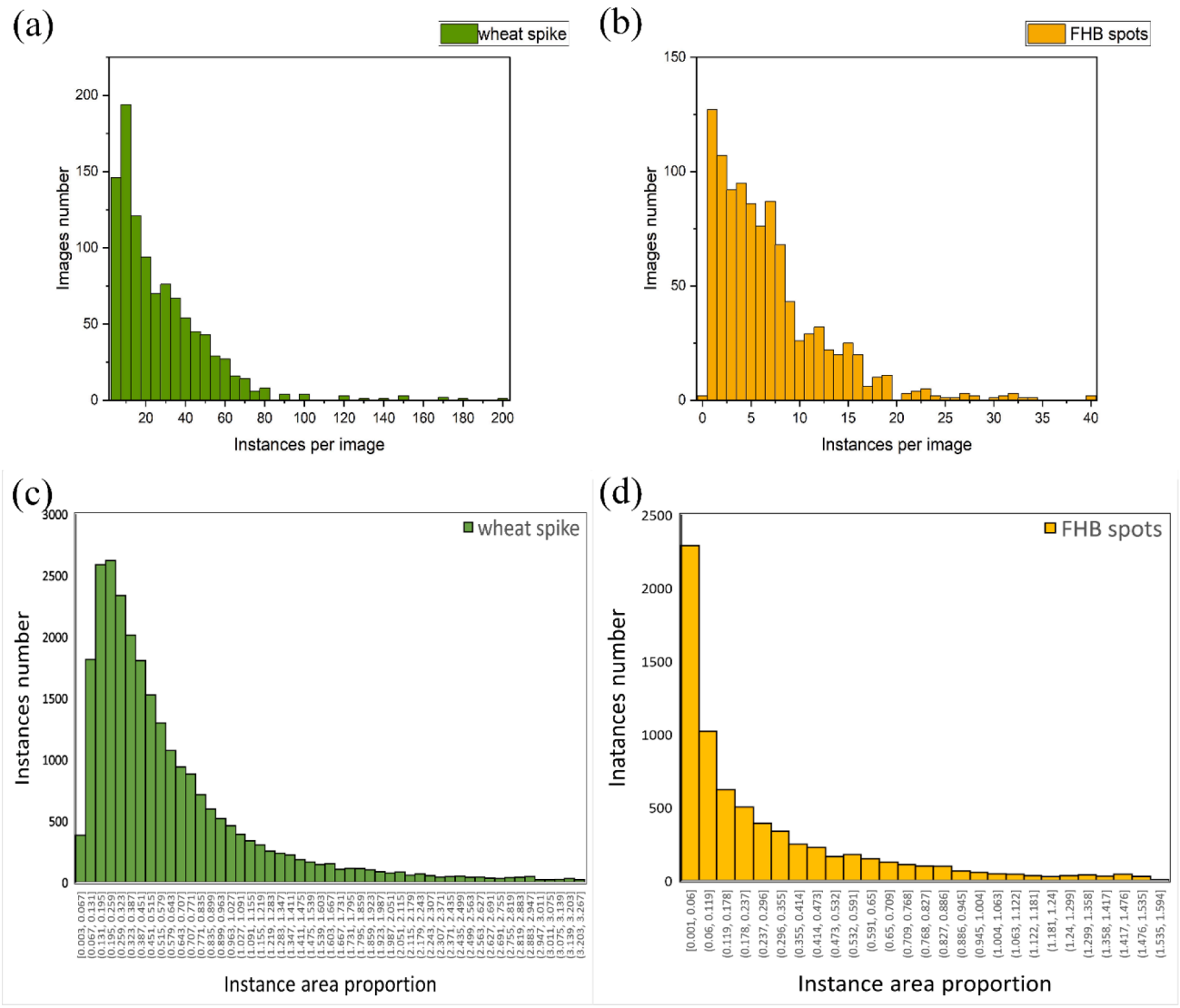

Fig. 3. Some statistic of dataset. (a) number of wheat spike mask for all images (b) number of FHB spots mask for all images (c) mask area proportion histogram of wheat spike (d) mask area proportion histogram of FHB spots.

2.2. Overview of the DeepFHB model

The dense wheat spike and the area of Fusarium head blight can be refined and segmented by our DeepFHB model. The architecture of the proposed network is illustrated in Fig. 4. It consists of three modules: an object detection module, a lightweight incoherent region detector, and a refinement transformer. In the object detection module, we predict bounding boxes as region proposals and produce an initial mask prediction at the low-resolution level. The incoherent region detector, which is known for its lightweight design, processes a rough initial mask in conjunction with multi-scale features to identify incoherent regions across various scale. Moreover, the refinement transformer uses the incoherent points identified on the constructed quadtree as input for the final segmentation refinement. The workflow of the proposed DeepFHB network is as follows,

1. Images are fed to the object detection module, containing the backbone network and FPN(Lin et al., 2017a) network, with which bounding boxes and an initial course mask at low-resolution are predicted.
2. A pyramid is constructed to identify incoherent regions in multiple scale. We employ a lightweight detector and organize nodes with quadtree structure. Given that only a fraction of high-resolution image features needs to be processed by the refinement network, this allows our network to save huge memory and computational burdens.
3. The refinement transformer was designed for predicting highly accurate instance segmentation masks. Different from the traditional transformer in the encoder part, the encoder here is mainly composed of two parts, the node encoder and sequence encoder. The transformer performs both global spatial and inter-scale reasoning. The Decoder part is a two-layer small MLP, which can decode the output query label of each node in the quadtree, in order to predict the final mask labels.
4. During the training process, the entire DeepFHB framework is designed to be trained end-to-end. This approach was applied to achieve refined segmentation of both wheat spikes and areas affected by Fusarium Head Blight (FHB) in the test images.

2.3. Object detection module

The object detection module employs a network architecture similar to ResNet, augmented with deformable convolutional layers and group normalization. The deformable convolutions (Dai et al., 2017) enable

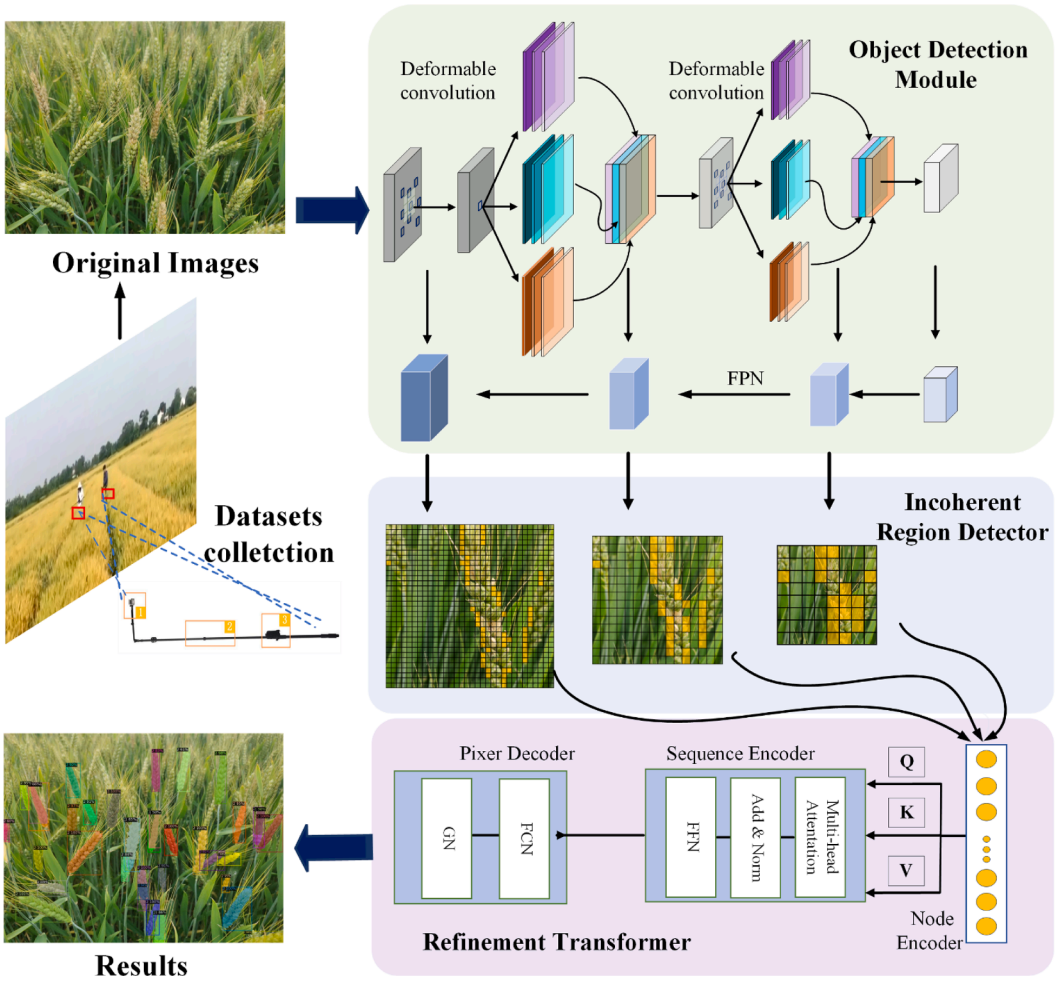

Fig. 4. The framework of DeepFHB.

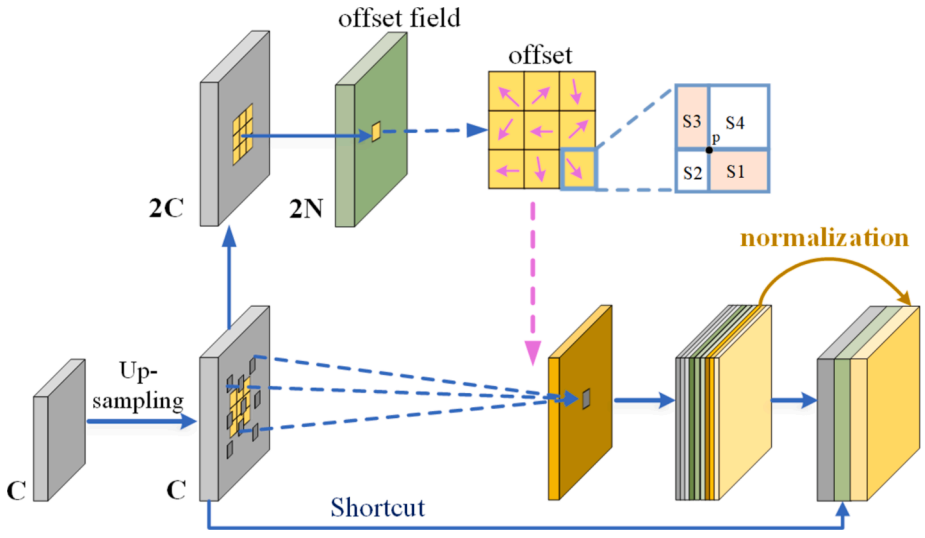

Fig. 5. Structure of Normalized deformable convolution.

the sub-network to adapt to the spatial characteristics of irregularly shaped objects like wheat spikes, while group normalization (Wu and He, 2018) addresses the dependency of batch normalization on batch size. Convolution kernels are crucial in neural networks for extracting object features. However, due to the non-rigid and highly variable nature of wheat spike boundaries, deformable convolutions are favored over standard convolutions. Traditional convolution kernels, which are typically rectangular, are optimized for extracting features from objects with stable and fixed shapes. Their effectiveness diminishes when dealing with objects that do not conform to fixed geometric patterns,

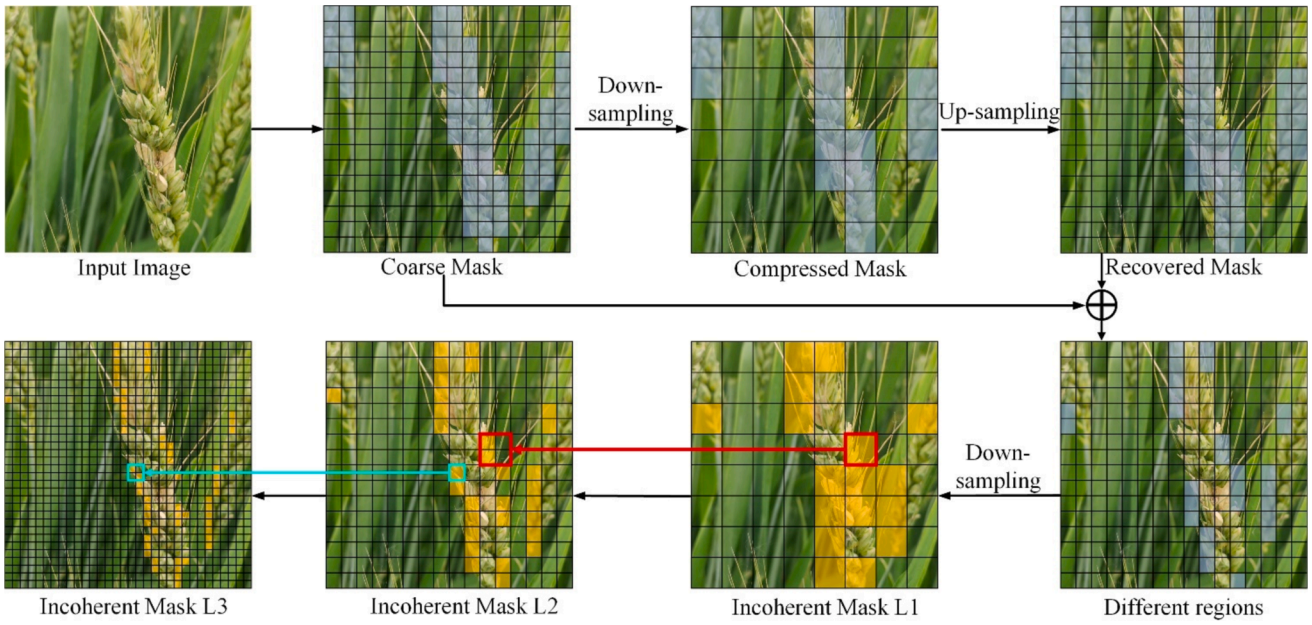

Fig. 6. Quadtree on Incoherent Regions.

thus reducing their generalization capability.Given the diverse shooting angles encountered during close-up or distant capturing, along with significant variations in the shapes and scales of wheat spikes, and the dense distribution in real field environments often resulting in severe occlusions, traditional convolutional approaches face significant challenges. Deformable convolutions offer a solution by introducing an offset to each sampling point in the kernel, thereby allowing the network to sample from regions adjacent to the original points and transcend the limitations of traditional rectangular sampling areas. This adaptation not only broadens the receptive field but also significantly enhances the ability to capture and utilize features from deformable objects, making the feature extraction process more effective for complex shapes encountered in agricultural settings. Ability of the deformable wheat spikes.

2.3.1. The deformable convolution

Fig. 5 exhibits the normalized deformable convolution kernel, where the arrow visually illustrates the augmented offset. The initial point of the arrow corresponds to the sampling location of the conventional convolution kernel, while the terminus represents the updated sampling location post-offset. Notably, the offset sampling points are evident as irregular shapes rather than rectangular forms, as depicted by the scattered dots. In the traditional convolution framework, given an input feature map x , the value of each pixel p in the output feature map y is computed as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot (p_0 + p_n) \tag{1}$$

here, p_0 represents the central position of the local receptive field R, while p_n denotes the relative position within the receptive field around p_0. Moreover, $w(p_n)$ symbolizes a learnable parameter that contributes to the convolution operation. The deformable convolution modifies this framework by adding an offset Δp_n to the coordinates of each sampling point to realize the coordinate offset of the sampling point. The definition of dconv can be represented as:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot (p_0 + p_n + \Delta p_n) \tag{2}$$

Compared with ordinary convolution, deformable convolution can adjust the range of convolution operation by a learnable parameter Δp_n. Due to the existence of Δp_n, the sampling points spread into a non-grid shape. Since the value of Δp_n may be a fractional value, bilinear interpolation is used to calculate $x(p_0 + p_n + \Delta p_n)$.

2.3.2. Group normalization

In image segmentation, the batch size is usually set to a smaller value to save GPU memory. However, the consequence of a small batch size may lead to inaccurate calculations of the mean and variance, thereby reducing the performance of BN (Batch Normalization). To address this issue, we choose Group Normalization (GN) instead of BN. GN enhances the model's normalization capability even when the batch size is very small. When GN calculates the mean and standard deviation, the channel dimension of each feature map is divided into G groups, with each group containing C/G channels. It then computes the average and standard deviation of the pixels within these channels. Each group of channels is normalized independently using its corresponding parameters, thus GN's operation is not affected by batch size and the process is more stable than BN. The derivation process is illustrated as follows:

$$\widehat{x}_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3}$$

where x denotes the tensor calculated by the feature map, and i presents the index number. The normalized tensor x^i is described as $x^i = [x_{iN}, x_{iC}, x_{iH}, x_{iB}]$. μ and σ are the mean and standard deviation respectively, which can be formulated as:

$$\mu_i(x) = \frac{1}{(C/G)HW} \sum_{c=gC/G}^{(g+1)C/G} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{nchw} \tag{4}$$

$$\sigma_i(x) = \sqrt{\frac{1}{(C/G)HW} \sum_{c=gC/G}^{(g+1)C/G} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_{nchw} - \mu_i(x))^2 + \varepsilon} \tag{5}$$

where ε is a very small constant to ensure that $\sigma \geq 0$. n, c, g, h, and w are index numbers; B, C, G, H, and W are value ranges, where G is the artificially set number of groups, C / G is the number of channels per group.

2.4. Incoherent region detector

We propose an incoherent region detector for detecting error-prone

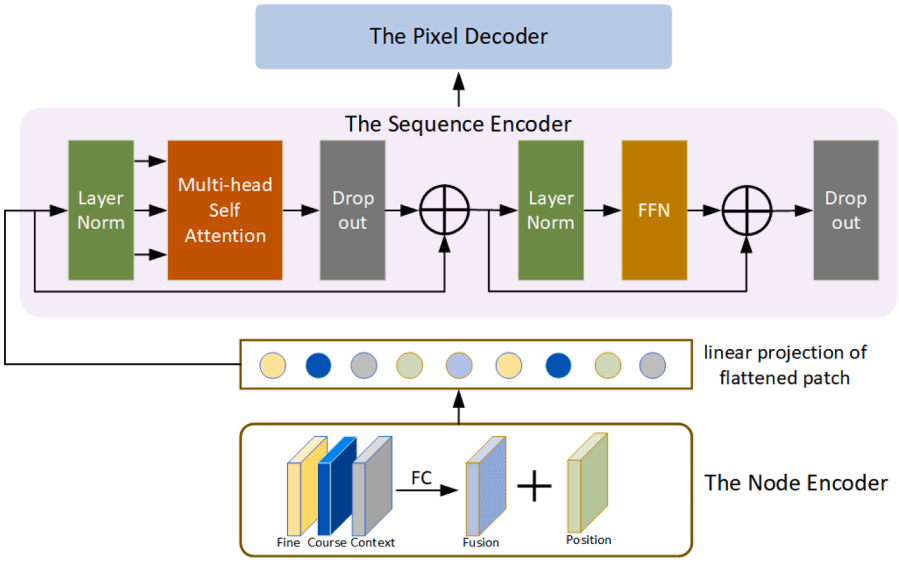

Fig. 7. The architecture of Refinement Transformer.

regions where mask information is lost due to reduced spatial resolution. We compute it by down sampling compression and up-sampling recovering. As illustrated in Fig. 6, the up-sampling operation fails to recover the course segmentation mask. Due to its properties, a large part of the prediction error is concentrated in the incoherent regions. We computer residuals on regions where mask information is lost between the coarse mask and recovered mask. The incoherent region R_l of scale L can be formulated as follows:

$$R_L = f(A_{L-1} \oplus U(D(A_{L-1}))) \tag{6}$$

here, A_{L-1} is the binary ground-truth instance mask for objects at scale level L-1. The resolution of each scale level differs by a factor of 2. The nearest neighbor down-sampling and up-sampling are denoted by D and U, respectively. $\oplus$ indicates a logical "exclusive OR" operation, and f represents a 2 $\times$ down-sampling by performing a logical "OR" operation in each 2 $\times$ 2 neighborhood.

Given an input image, using its ground truth coarse mask, we can calculate the first level of the incoherent mask by a forementioned process, and then we can further decompose the quadtree nodes from Level1 to Level2. From Level2 to Level3, we can further decompose and construct a quadtree for three levels of incoherent nodes.

2.5. Refinement transformer

In contrast to standard transformer encoder, the encoder of refinement transformer has two parts: the node encoder and the sequence encoder. The input node sequence consists of three different levels of incoherent region nodes from the quadtree. The size of the sequence is C $\times$ N, where C is the dimension of the feature channel and N is the total number of nodes. To enrich the feature of incoherent points, the node encoder encodes each node of the quadtree with the following four information: 1) Fine-grained features extracted from the FPN of the current level. 2) Semantic information provided from the initial coarse mask prediction region. 3) Relationship and distance information between nodes (encapsulated by relative position encoding ROI). 4) context information and self-information of each node. Refinement Transformer module can capture global information and abundant contextual information.

2.5.1. The sequence encoder

The sequence encoder is composed of transformer encoder modules stacked three times, which contains two sub-layers, a multi-head attention layer and a fully connected feed-forward neural network referred to as an MLP. The structure is illustrated in Fig. 4. For the nth layer of the sequence encoder module, with x_n denoting its input and x_{n-1} representing its output, the entire computation process can be described as follows:

$$x_n = FFN\left[LN\left(x_n^{'}\right)\right] + x_n^{'}, n = 1, 2, 3 \tag{7}$$

$$x_n^{'} = MSA[LN(x_{n-1})] + x_{n-1}, n = 1, 2, 3 \tag{8}$$

In the equation, MSA refers to Multi-head Self Attention illustrated in Fig. 5. FFN stands for Feed-Forward Neural network. LN represents layer normalization, which corresponds to the Layer Norm in Fig. 7.

Residual connections are employed between these sub-layers. Layer normalization and dropout layers are incorporated in the network architecture to facilitate improved convergence and mitigate overfitting. The utilization of multi-head attention enables the current node to attend not only to the current pixels but also to acquire contextual semantics. Transformer encoder blocks greatly enhance the capability to capture diverse local information and leverage the self-attention mechanism to explore the potential of feature representation. Evaluation on the FHB-SA dataset demonstrates the superior performance of transformer encoder blocks in effectively detecting occluded objects with high-density. The Pixel Decoder is a small two-layer MLP (Multilayer Perceptron) that decodes each node's output query and predicts the final mask label.

2.5.2. Loss function

Based on the constructed quadtree, all irrelevant nodes detected across quadtree levels form a sequence for parallel prediction. During training, a multi-task loss is employed,

$$L = \lambda_{Dec} L_{Dec} + \lambda_{Crs} L_{Crs} + \lambda_{\mathrm{Ref}} L_{\mathrm{Ref}} + \lambda_{Inc} L_{Inc} \tag{9}$$

where the L_{Dec} denotes $\delta = f(I_b)$ the combined losses associated with bounding box position regression and classification, as determined by the box detection head in the decoder. It comprises a loss component for bounding box position regression, which measures the difference between the predicted and actual bounding box positions, and a classification loss, which evaluates the difference between the predicted and actual categories. The loss denoted as L_{crs} corresponds to the coarse-mask labels generated by the coarse-mask head, which can also serve as a representative loss value in classic instance segmentation methods like Mask R-CNN (He et al., 2017). Furthermore, L_{Ref} and L_{Inc} represent

Table 4
Hyperparameter Settings for DeepFHB model.

Hyperparameter	Values
Initial learning rate	0.005
Momentum	0.9
Warmup_iterations	1000
Max_iterations	9000
Warmup_ratio	0.001
Batch size	2
Number of classes	2

the losses associated with the predicted labels for incoherent nodes and the detection of incoherent regions, respectively. In this paper, the hyper-parameter values of $\lambda_{Dec}, \lambda_{Crs}, \lambda_{\mathrm{Ref}}, \lambda_{Inc}$ are set to$\{1.0, 1.0, 1.0, 0.5\}$.

3. Experimental results and analysis

In Section 3, using the specially constructed dataset FHB-SA, we conduct extensive experiments to validate the effectiveness of our presented methods. This rigorous evaluation not only benchmarks against several state-of-the-art methods but also clarifies the practical implications of our research. Initially, in Section 3.1, we detail the baseline comparisons, evaluation metrics, and experimental configurations employed to ensure comprehensive testing. This setup is crucial for demonstrating the efficacy of our approach and its reliability in comparison with existing standards. Section 3.2 reports the results of ablation studies, shedding light on the significance of each component in our model. Subsequently, in section 3.3, a comparative analysis with state-of-the-art methods is presented, highlighting the enhancements and positioning of our model within the broader research landscape. Finally, Section 3.4 provides a qualitative assessment of our findings, illustrating the visual effectiveness and practical applications of our approach.e baseline, evaluation metrics, and experimental configurations are elaborated upon in Section 3.1. The comparative analysis with state-of-the-art methods is presented in Section 3.2. The results of the ablation experiments are reported in Section 3.3, while the qualitative results are provided in Section 3.4.

3.1. Experimental settings

3.1.1. Baseline

The baseline model for comparison consists of both Mask R-CNN(He et al., 2017) and Cascade Mask R-CNN(Cai and Vasconcelos, 2018) with Feature Pyramid Network (FPN)(Lin et al., 2017a). Mask R-CNN, short

Table 5
Ablation study on the major components in detection pipeline on FHB-SA dataset.

Method	Backbone	Refinement Transformer	DConv	GN	mAP^{Box}	AP^{Box}_{50}	AP^{Box}_{75}
Mask R-CNN	ResNet50-FPN				37.271	62.11	39.329
		✓			37.308	63.182	39.443
		✓	✓		38.035	64.126	40.542
		✓		✓	37.559	63.976	39.842
		✓	✓	✓	**38.76**	**64.408**	**41.769**
	ResNet101-FPN				36.985	61.4	39.523
		✓			37.906	63.274	40
		✓	✓		37.991	63.771	40.523
		✓		✓	38.521	63.576	41.589
		✓	✓	✓	**38.621**	**64.057**	**41.809**
Cascade Mask R-CNN	ResNet50-FPN				39.058	61.842	42.024
		✓			39.596	62.808	42.442
		✓	✓		39.986	63.357	43.387
		✓		✓	40.512	63.145	44.409
		✓	✓	✓	**40.958**	**64.322**	**44.635**
	ResNet101-FPN				37.677	60.985	39.844
		✓			37.906	63.274	40
		✓	✓		37.991	63.771	40.523
		✓		✓	38.521	63.576	41.589
		✓	✓	✓	**38.612**	**64.057**	**41.809**

Table 6
Ablation study on the major components in segmentation pipeline on FHB-SA dataset.

Method	Backbone	Refinement Transformer	DConv	GN	mAP^{Mask}	AP^{Mask}_{50}	AP^{Mask}_{75}
Mask R-CNN	ResNet50-FPN				36.495	62.431	39.832
		✓			36.66	62.726	40.209
		✓	✓		37.555	**64.245**	40.925
		✓		✓	38.306	64.69	42.534
		✓	✓	✓	**38.429**	**64.71**	**42.752**
	ResNet101-FPN				36.415	61.796	40.472
		✓			38.283	**64.327**	42.329
		✓	✓		38.236	64.436	42.589
		✓		✓	38.397	64.751	42.678
		✓	✓	✓	38.656	**64.966**	42.694
Cascade Mask R-CNN	ResNet50-FPN				36.694	62.084	40.63
		✓			38.045	63.162	42.287
		✓	✓		38.338	63.368	42.963
		✓		✓	38.503	63.441	42.997
		✓	✓	✓	38.601	63.612	43.094
	ResNet101-FPN				36.266	60.644	40.913
		✓			37.694	62.871	41.985
		✓	✓		38.358	63.388	43.25
		✓		✓	38.453	63.789	43.095
		✓	✓	✓	38.624	64.023	43.351

Table 7
Detection results on FHB-SA dataset. Faster R-CNN, Mask R-CNN and Cascade Mask R-CNN are the representative models of two-stage. FCOS and RetinaNet are the representative models of one-stage. PointRend is the representative models of query.

Methods	Backbone	mAP^{Box}	AP^{Box}_{50}	AP^{Box}_{75}	AP^{Box}_{s}	AP^{Box}_{m}	AP^{Box}_{l}
Faster R-CNN	ResNet50	37.513	62.895	40.066	3.288	29.546	46.889
Faster R-CNN	ResNet101	34.988	60.049	37.117	3.022	26.052	45.285
Mask R-CNN	ResNet50	37.271	62.11	39.329	2.603	29.633	46.501
Mask R-CNN	ResNet101	36.985	61.4	39.523	3.384	28.855	45.988
Cascade Mask R-CNN	ResNet50	39.058	61.842	42.024	4.389	30.75	48.282
Cascade Mask R-CNN	ResNet101	37.677	60.985	39.844	5.265	30.344	47.621
FCOS	ResNet50	34.744	61.023	35.896	1.924	27.722	44.514
MS R-CNN	ResNet50	37.971	62.51	41.229	6.703	29.733	47.301
HTC	ResNet50	38.871	61.61	42.329	5.603	30.933	47.901
RetinaNet	ResNet50	32.362	57.107	33.24	0.795	24.212	42.174
RetinaNet	ResNet101	32.541	57.725	33.297	1.342	23.767	42.998
PointRend	ResNet50	37.2	61.728	39.401	3.253	29.351	46.622
Our method + Mas	ResNet50	38.76	**64.408**	41.769	**8.045**	32.376	47.679
Our method + Mas	ResNet101	38.612	64.057	41.809	7.707	32.106	46.437
Our method + Cas	ResNet50	**39.958**	63.322	**43.235**	5.132	**32.482**	**48.744**
Our method + Cas	ResNet101	39.401	62.284	42.559	6.768	31.575	48.035

Table 8
Instance segmentation results on FHB-SA dataset. Mask R-CNN, Cascade Mask R-CNN and HTC are the representative models of two-stage. PointRend is the representative models of query.

Methods	Backbone	mAP^{Mask}	AP^{Mask}_{50}	AP^{Mask}_{75}	AP^{Mask}_{s}	AP^{Mask}_{m}	AP^{Mask}_{l}
Mask R-CNN	ResNet50	36.495	62.431	39.832	1.698	26.726	48.485
Mask R-CNN	ResNet101	36.415	61.796	40.472	1.756	26.465	48.103
Cascade Mask R-CNN	ResNet50	36.694	62.084	40.63	1.983	26.936	47.97
Cascade Mask R-CNN	ResNet101	36.266	60.644	40.913	1.885	26.442	47.197
MS R-CNN	ResNet50	37.695	62.931	42.532	1.798	27.226	49.085
HTC	ResNet50	37.695	62.631	41.332	1.798	27.326	49.185
PointRend	ResNet50	37.19	62.562	41.657	1.47	26.889	49.735
Our method + Mas	ResNet50	**38.429**	64.71	**42.752**	2.49	**30.129**	**50.949**
Our method + Mas	ResNet101	38.138	**64.966**	42.329	**2.568**	29.925	49.732
Our method + Cas	ResNet50	38.045	63.368	42.287	2.088	28.434	50.882
Our method + Cas	ResNet101	37.694	62.871	41.985	2.366	28.627	50.779

Table 9
Comparison of AP Value: Our Method vs. Other Models.

Methods	Backbone	AP^{Box}_{50}	AP^{Mask}_{50}	AP^{Box}_{75}	AP^{Mask}_{75}
Su et al. *(2020)	ResNet101	56.1	55.853	33.059-	32.332
Zhang et al.* (2022)	ResNet50	63.187	63.736	39.927	40.192
Gao et al.* (2022)	ResNet50	57.12	57.212	37.06	38.576
Our method	ResNet50	64.408	64.71	41.769	42.752

for Mask Region Convolutional Neural Network, is an advanced deep learning algorithm within the domain of computer vision designed specifically for addressing the challenge of instance segmentation. Cascade Mask R-CNN, incorporates two intermediary stages dedicated to object detection and mask segmentation within the processing pipeline. These stages are inserted between the feature extraction phase and the final object detection (instance segmentation) heads of the Faster R-CNN(Ren et al., 2015) (or Mask R-CNN) framework. This augmentation of intermediate stages is what distinguishes Cascade R-CNN(Cai and Vasconcelos, 2018) from its predecessors, providing it with enhanced performance and refinement in object detection and mask segmentation tasks. Notably, this algorithm has achieved notable success in applications such as the identification of plant diseases and fruits, underscoring its effectiveness in complex visual recognition tasks.

3.1.2. Evaluation metrics

In the conducted experiments, we employ the evaluation metrics as defined in the context of the COCO dataset. The metric of Average

Table 10
Comparison of detection results of different models on AR.

Methods	Backbone	$AR1^{Box}$	$AR10^{Box}$	$AR100^{Box}$	$AR100^{Box}_{s}$	$AR100^{Box}_{m}$	$AR100^{Box}_{l}$
Faster R-CNN	ResNet50	4.6	24.7	47.6	7.4	40.6	56.7
Faster R-CNN	ResNet101	4.6	23.7	44.4	7.6	37.1	55.3
Mask R-CNN	ResNet50	4.6	24.4	46.9	6.6	40	55.7
Mask R-CNN	ResNet101	4.7	23.8	46.3	6.8	39.7	54.8
Cascade Mask R-CNN	ResNet50	4.7	24.8	49.2	8.8	42.4	57.9
Cascade Mask R-CNN	ResNet101	4.7	24.7	48.3	9.9	41.3	57.0
FCOS	ResNet50	3.9	23.5	46.5	3.1	40.4	57.6
RetinaNet	ResNet50	3.9	21.7	45.1	3.9	36.8	57
RetinaNet	ResNet101	4.1	21.8	44.9	5.2	35.7	57.1
PointRend	ResNet50	4.5	24.2	47.2	7.6	40.1	56.3
Our method + Mas	ResNet50	**4.8**	25.2	48.9	7.9	42.4	57.9
Our method + Mas	ResNet101	4.6	24.8	48.8	8.1	43.5	56.2
Our method + Cas	ResNet50	4.7	25.5	50.7	9.3	**44.6**	59.7
Our method + Cas	ResNet101	4.8	**26.2**	**51.9**	**12.2**	44.4	**62.6**

Table 11
Comparison of instance segmentation results of different models on AR.

Methods	Backbone	$AR1^{Mask}$	$AR10^{Mask}$	$AR100^{Mask}$	$AR100_s^{Mask}$	$AR100_m^{Mask}$	$AR100_l^{Mask}$
Mask R-CNN	ResNet50	4.6	23.9	45.4	6.6	39.1	54.0
Mask R-CNN	ResNet101	4.6	23.5	44.9	7.0	39.0	53.1
Cascade Mask R-CNN	ResNet50	4.4	23.5	45.8	8.5	39.7	53.9
Cascade Mask R-CNN	ResNet101	4.5	23.1	44.7	8.7	38.7	52.8
PointRend	ResNet50	4.6	23.9	46.3	7.8	39.6	55.4
Our method + Mas	ResNet50	**4.8**	**25.3**	47.9	8.6	41.7	57.3
Our method + Mas	ResNet101	4.7	24.6	47.3	9.5	41.7	55.4
Our method + Cas	ResNet50	4.6	24.3	48.5	9.8	**42.1**	55.9
Our method + Cas	ResNet101	4.7	25	**48.6**	**12.1**	41.8	**58.5**

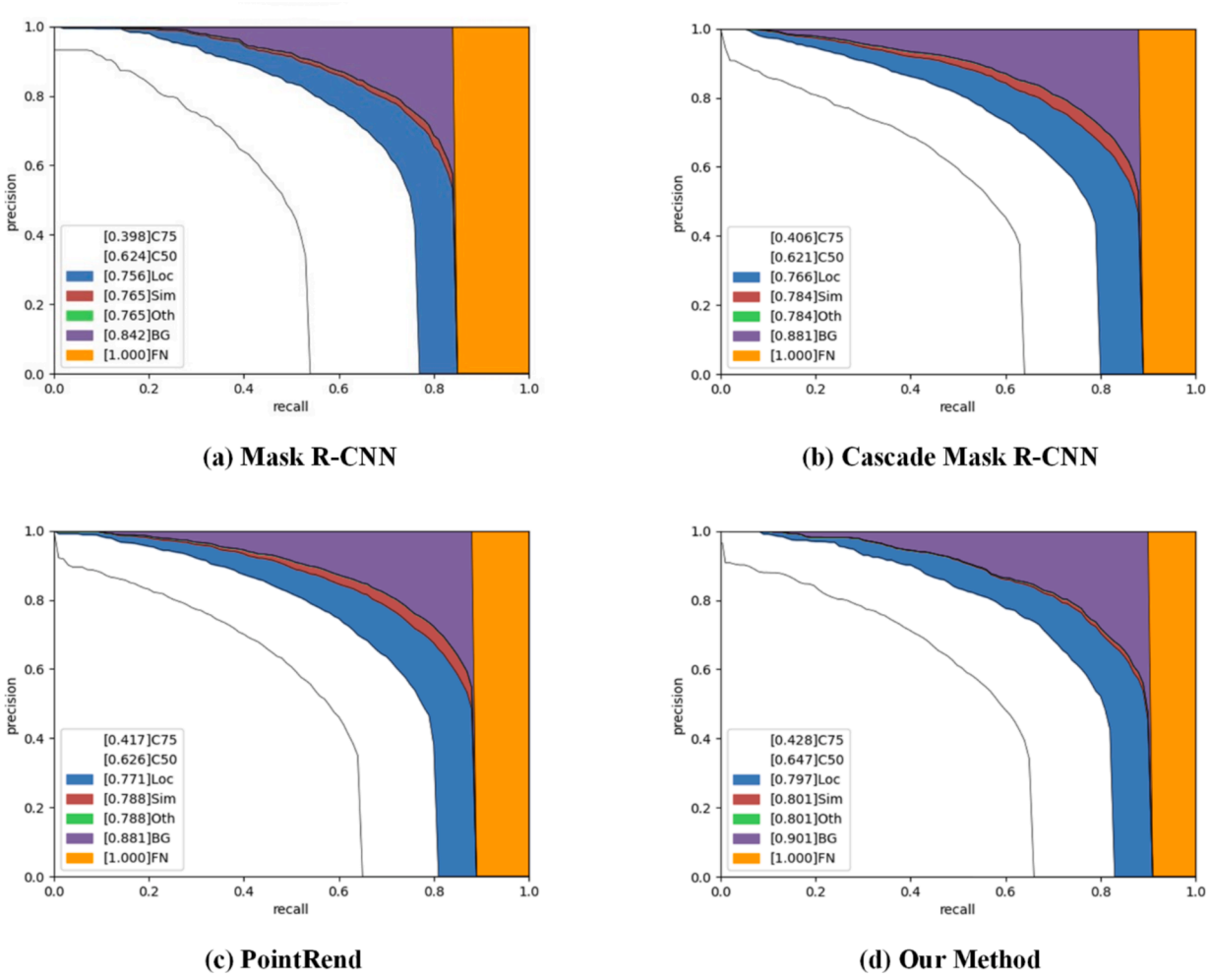

Fig. 8. P-R curve. Quantitative evaluation of segmentation performance compared to Mask R-CNN, Cascade Mask R-CNN and PointRend. The baseline is performed by Mask R-CNN with ResNet-50, and our method is also deployed in this approach.

precision (AP) is conceptualized as the area under the precision-recall curve. The specific definition of AP is given by Formula 10–13. AP at IoU (Intersection over Union) equal to 0.5 and 0.75 are denoted as AP@50 (AP_{50}) and AP@75 (AP_{75}) respectively, while mean AP (mAP) is calculated for IoU thresholds ranging from 0.5 to 0.95(inclusive), incremented by 0.05.

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$AP = \int_0^1 precision(recall)d(recall) \tag{12}$$

$$mAP = \frac{1}{N}\sum_{j=1}^{N} AP_j \tag{13}$$

where TP (True Positive), FP (False Positive), FN (False Negative) signify the quantities of correctly identified, incorrectly identified, and undetected wheat spikes, respectively. Simultaneously, the metrics AP_s, AP_m, AP_l defined within the COCO dataset are utilized in our research. These metrics quantify the detection accuracy for target sizes of varying scales. However, due to the wheat spike in the dataset constituting a

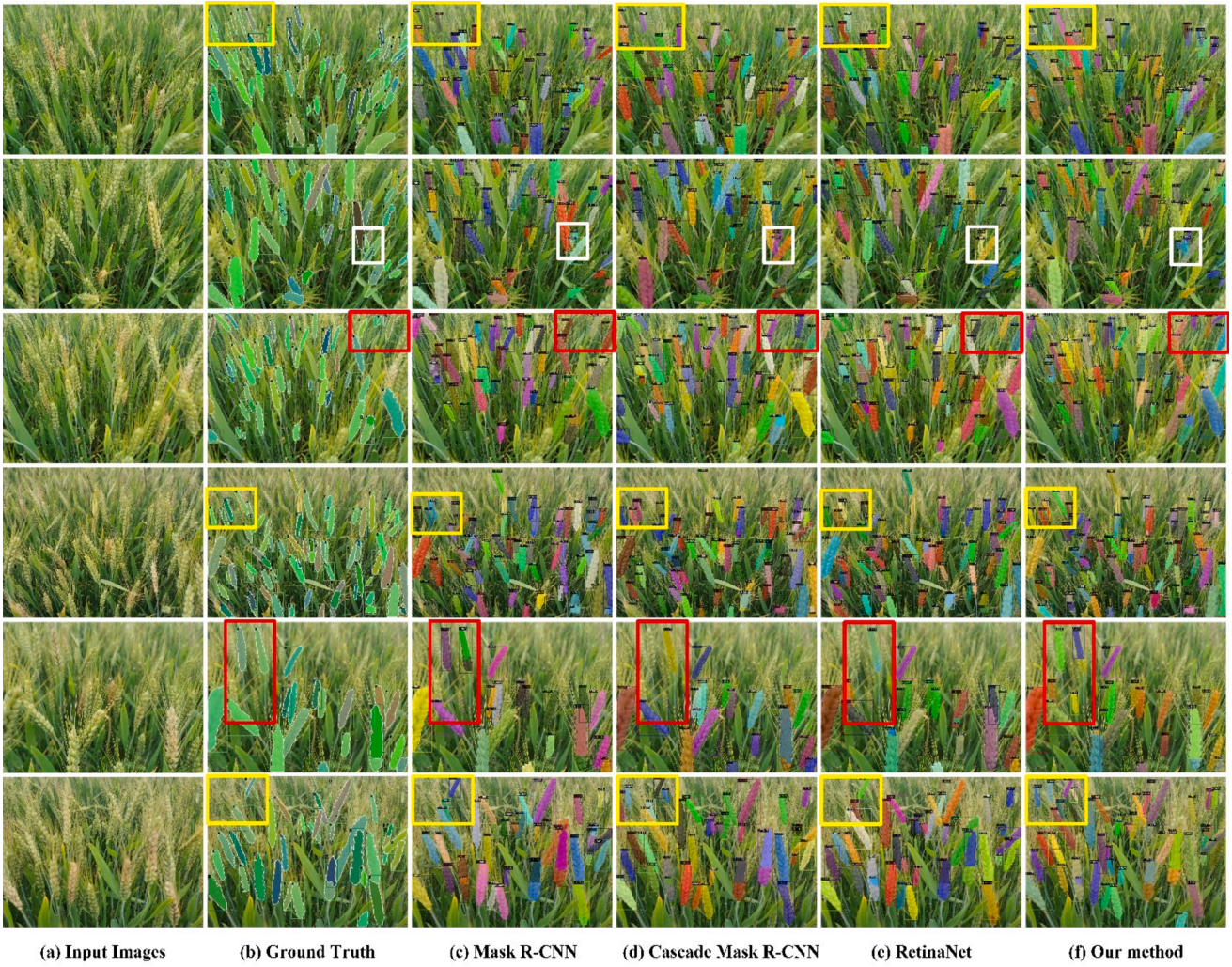

Fig. 9. Qualitative segmentation performance on FHB-SA dataset. From left to right column: input images, performance comparison on ground truth, Mask R-CNN, Cascade Mask R-CNN, RetinaNet and our method model. The red, white and yellow boxes highlight some representative comparisons.

significant proportion of the image, we restrict our application to AP_m (pertaining to medium targets) and AP_l (relevant to large targets) for the purpose of evaluation. It should be noted that the AP metric is extensively leveraged in the domain of object detection, offering a comprehensive measure of a model's detection and segmentation performance.

3.1.3. Experiment platform

The experiments were conducted on a hardware platform with an Intel Core i9-9900 k CPU, 128 GB RAM, and an NVIDIA TITAN RTX GPU (24 GB memory). We trained the instance segmentation network on Ubuntu 18.04 with python3.7 and CUDA 9.1 in the PyTorch deep learning framework. We set the learning rate to 0.005 in the initial state, training for a total of 9000 iterations. The weight decay method was linear, and the weight factor was 0.001. We trained all the models by end-to-end. The hyperparameters of the model are shown in Table 4.

3.2. Ablation study

The proposed segmentation detector for Fusarium Head Blight (FHB) has introduced three key components, namely the refinement transformer module, DConv Nets (deformable convolution), and the GN method (Group Normalization). An ablation study is conducted to dissect the individual contributions of each component, with the baseline model being Mask R-CNN with FPN and Cascade Mask R-CNN with FPN. These three components are systematically integrated into the baseline model one by one, and the resulting detection and segmentation outcomes are presented in Table 5 and Table 6, respectively. Initially, we introduce the refinement transformer module to the baseline, as depicted in the second row of Table 5. When the ResNet101 backbone is employed, the refinement transformer module exhibits a notable improvement of 2.531 % in AP50 for segmentation. Moving forward, the third row of Table 6 showcases the impact of DConv Nets, which elevate the performance from 62.726 % to 64.245 %. This improvement arises from the ability of DConv Nets to diversify the training samples, thereby benefiting the recognition of FHB disease. Finally, the GN method further enhances performance, elevating it from 64.245 % to 64.71 %, as evidenced in the fifth row of Table 6. This improvement results from the enhancement of feature representation achieved by the group normalization instead of batch normalization.

3.3. Comparison with State-of-the-Arts

In this section, we undertake a comparative study to validate the efficiency and effectiveness of the proposed framework. Our objective is to assess the performance of our model in the context of wheat disease detection and edge extraction in comparison to other detection and instance segmentation methods. We assessed the performance of our method for detecting wheat spikes and diseases by comparing it with eight detectors: Faster R-CNN(Ren et al., 2015), Mask R-CNN(He et al., 2017), Cascade Mask R-CNN(Cai and Vasconcelos, 2018), FCOS(Tian

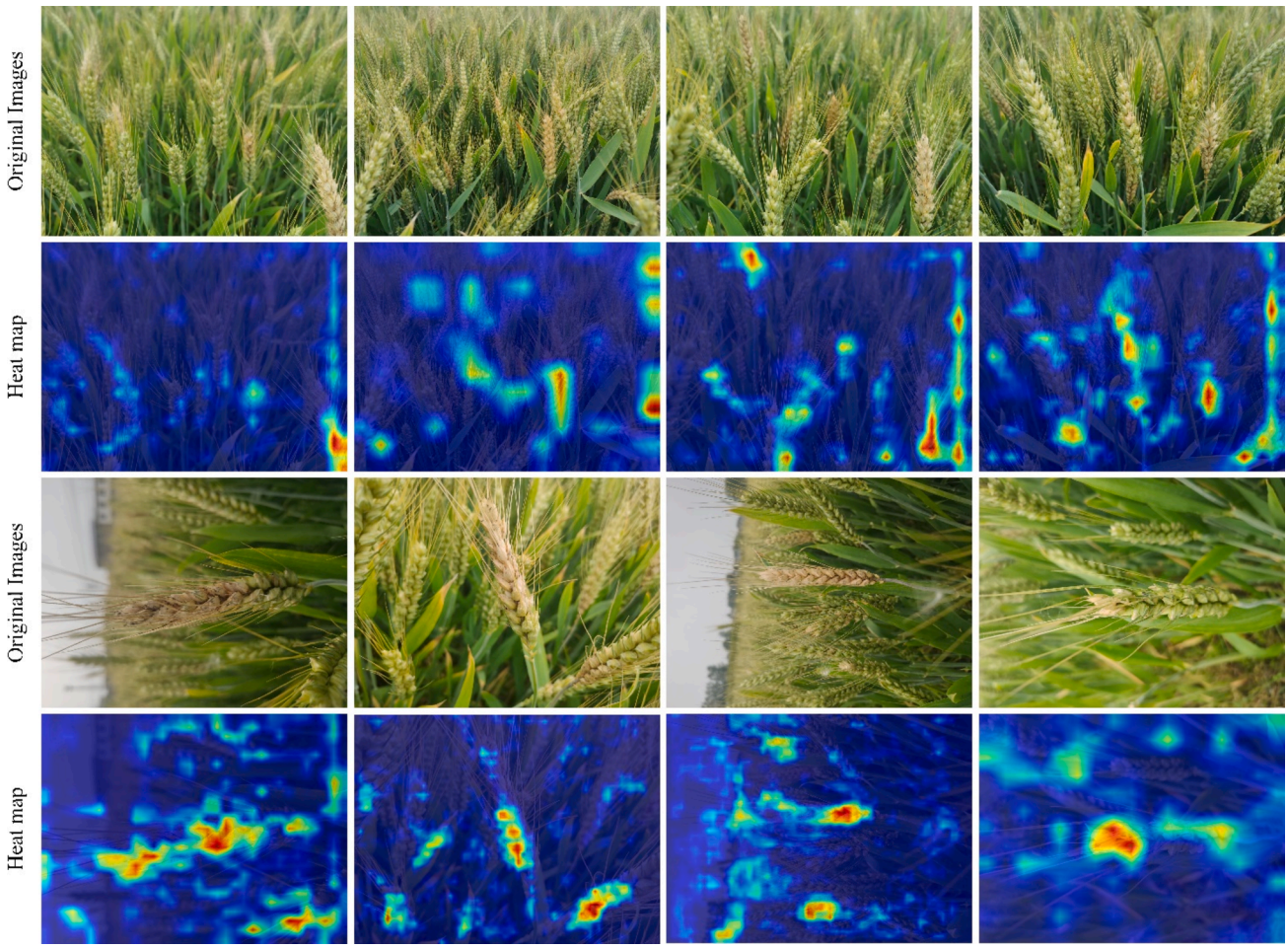

Fig. 10. Examples of visualization results. From the top to bottom: high-density wheat spikes in the field, the heat map of high-density wheat spikes in the field, sparse wheat spikes in the field, the heat map of sparse wheat spikes in the field. The red region represents the region of interest extracted by the model.

et al., 2019), MS R-CNN(Huang et al., 2019) , HTC(Chen et al., 2019), RetinaNet(Lin et al., 2017b) and PointRend(Kirillov et al., 2020). As shown in Table 7, out method outperforms other state-of-the-art detectors. The proposed method achieves 64.408 % AP50 on FHB-SA datasets, 7.301 % improvements of RetinaNet, 3.385 % improvements of FCOS and 2.68 % improvements of PointRend. Noticeably, its performance on small objects is 5.442 % AP higher than Mask R-CNN. We then compare the segmentation performance of our method with several other state-of-the-art methods on FHB-SA dataset. The results are listed in Table 8. Our method outperforms Mask R-CNN by 3.17 % when the backbone is ResNet101.

Furthermore, to assess our method's localization and segmentation performance, we conducted a comparative analysis against state-of-the-art techniques using recall metrics, as presented in Table 10 and Table 11. The tables clearly indicate that our approach excels in both localization and segmentation refinement. Specifically, our proposed method achieves an impressive 51.9 % Average Recall (AR), surpassing competing methods. Additionally, we investigated the accuracy of localizing and segmenting wheat spikes and Fusarium Head Blight (FHB) spots across various scales. For instance, our method achieves AR values of 12.2 %, 44.4 %, and 62.6 % for small, medium, and large agricultural FHB spots, respectively, outperforming alternative methods.

Table 9 presents a comprehensive comparison of our method against three other studies in the field, namely Su et al. (2020), Zhang et al. (2022b) and Gao et al. (2022). Our model outperforms in both bounding box (bbox) and mask segmentation compared to these existing methods. Due to the unavailability of datasets and code from these studies, we replicated their experiments using the algorithms described in their respective papers. This approach ensures that the comparisons are conducted under consistent and fair conditions, providing a reliable basis for evaluating the relative performance of our method.

3.4. Qualitative results visualization

Precision-recall (PR) curves. In terms of qualitative evaluation, we visualize some of detection results to further substantiate our approach. Fig. 8 provides an illustration of precision-recall (PR) curves for various metrics (C75, C50, Loc, Sim, Oth, BG, and FN) as applied in our experimental setup. The PR curves compare our model (d) with existing state-of-the-art methods: Mask R-CNN (a), Cascade Mask R-CNN (b), and PointRend (c), specifically within the heterochromatic category. These curves take recall as the axis and precision as the ordinate axis, with the performance of each model reflected by the area below the curve. The larger the area, the better the performance of the model. Our model demonstrates a notable performance advantage, as evidenced by a C50 score of 0.647, marking a substantial improvement of 2.3 % over Mask R-CNN, 2.6 % over Cascade Mask R-CNN, and 2.1 % over PointRend. Even under more stringent IoU thresholds (AP75), our method outperforms the baselines by a margin of 3 %, maintaining superior performance. Furthermore, by effectively reducing false positives caused by background interference, we have improved detection accuracy from 0.842 to 0.901. These enhancements are particularly visible in Fig. 8, where the area under our model's curve is visibly larger compared to the others, underscoring its capability to minimize missed detections and confirming the effectiveness of our proposed refinement transformer

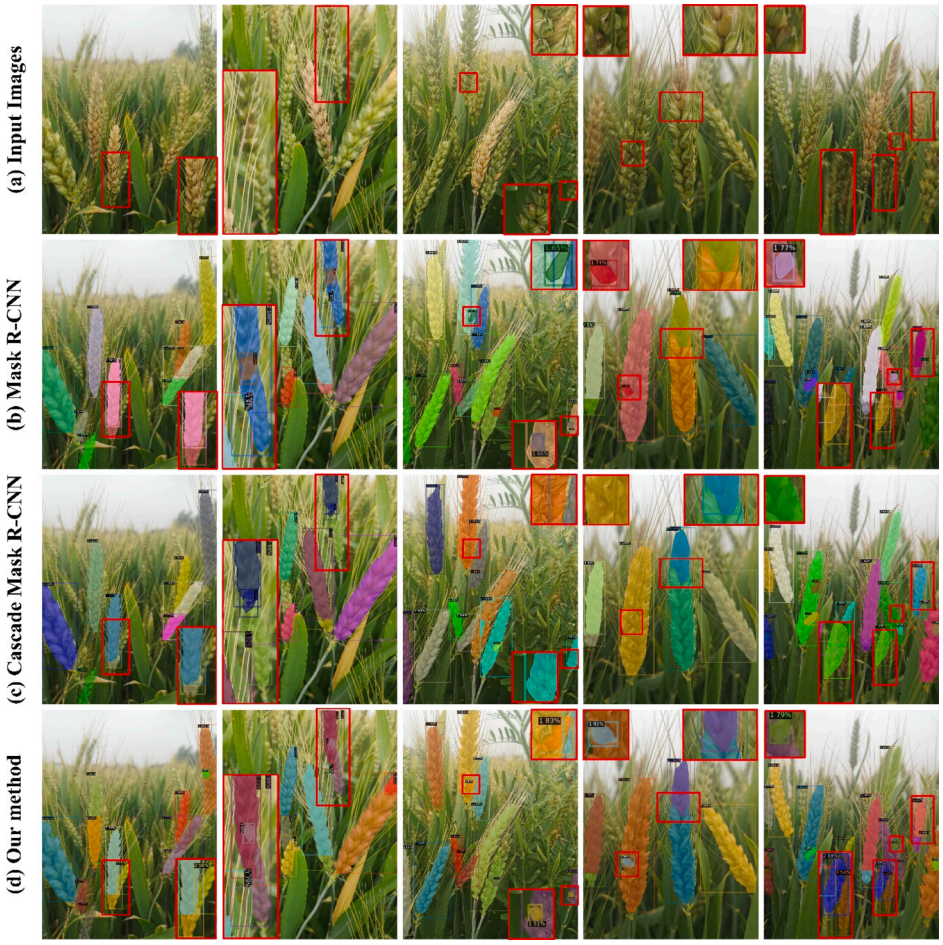

Fig. 11. Qualitatively comparisons of the low-density wheat spikes shot in scenes such as small targets, adjacent spikes, insufficient illumination of spikes, and wheat spikes being cut at image boundaries. From top to bottom are input images, and the results generated by Mask R-CNN, Cascade Mask R-CNN, and Our method. The red box part of each image has been enlarged to better show the experimental details.

module in enhancing two-stage detectors.

Qualitative Results. In order to ascertain the efficacy of our method, we compare our method with Mask R-CNN, Cascade Mask R-CNN and RetinaNet. Fig. 9 demonstrates several visualization results generated by various methods. It is evident from the visualizations that our method exhibited outstanding performance in both wheat spike detection and segmentation. This was manifested by a notable decrease in the occurrence of missed detections, thereby enhancing the overall accuracy of wheat spike segmentation. Especially, our method can precisely segment some instances without ground truth. The average visualization test time per image, with approximately 60 target instances, is 3 s. This performance aligns with the requirements for real-time detection and segmentation.

Visualization of heat maps. The intensity of the red color in the heat map, as illustrated in Fig. 10, is indicative of increased attention directed towards the features within that specific region. Upon comparing the disease images with their respective heat maps, it becomes apparent that the red areas align consistently with both the position and color of the disease spots in the images. This alignment suggests that the network effectively concentrates on the distinctive features associated with Fusarium Head Blight (FHB) diseases. The presented model adeptly directs its attention to the diseased spot areas of wheat spikes, effectively minimizing focus on extraneous and intricate backgrounds, thus obtaining higher disease identification accuracy than the other models.

Comparative Experiments and Discussions on the FHB-SA Dateset. In this study, we visualized the detection and segmentation results of the test dataset collected from real field scenarios. Figs. 11 and 12 illustrate the rfesults at varying densities of wheat spikes. To ensure a fair comparison with existing models such as Mask R-CNN and Cascade Mask R-CNN, we utilized PyTorch and Detectron2 (Wu et al., 2019) as the foundational deep learning frameworks. This approach ensures consistency and reliability in evaluating each model's performance in precision agricultural tasks within complex field environments. The segmentation results of our test images are extremely satisfactory. As depicted in Figs. 11 and 12, our DeepFHB model accurately segments each wheat spike, regardless of its sizes, orientations, number of spikes per image, locations, occlusions, and other characteristics.

We can observe that our model outperforms the other two comparative methods with complex backgrounds, dense spikes (Fig. 12), overlapping and occluding spikes, backlit collection (Fig. 11, fifth column), incomplete spikes at image edges, and small target lesions. Fig. 13 prominently displays the detection results in various occlusion scenarios. Taking the 1st, 2nd, and 5th columns of Fig. 11 and the 1st, 2nd, 4th, and 6th rows of Fig. 12 as examples, we observe the challenges in detecting overlapping wheat spikes, especially in the background obscured by wheat awns. Other methods either detected the overlapping wheat spikes as a single entity (as shown in the 2nd column of Fig. 11 and the 4th row of Fig. 12) or failed to detect the targets obscured by the awns (as in the 1st and 5th columns of Fig. 11, and the 2nd row of Fig. 12).

Our method not only successfully distinguishes individual overlapping wheat spikes but also identifies Fusarium Head Blight (FHB) lesions on the wheat spikes. Furthermore, in the last row of Fig. 12,

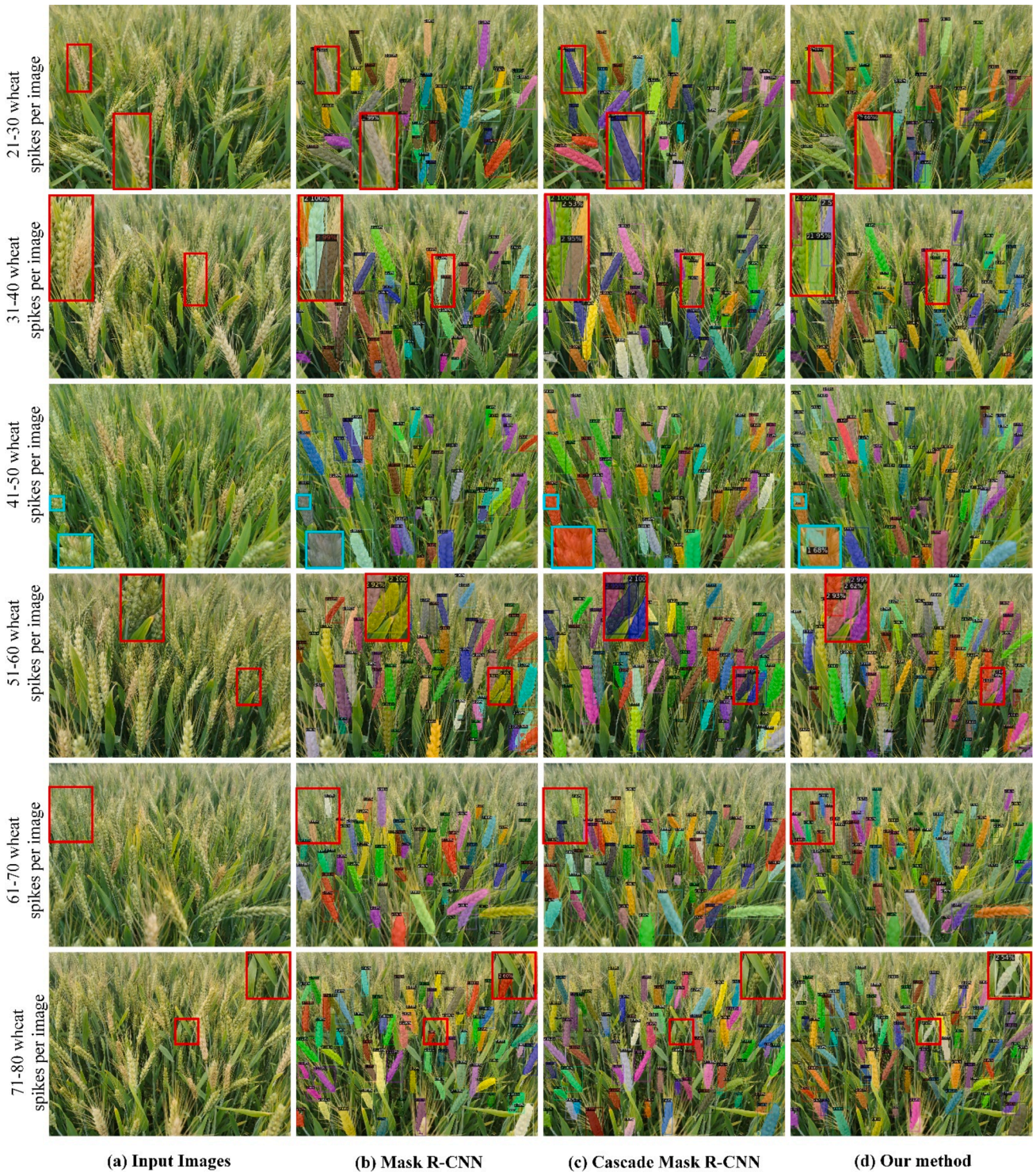

Fig. 12. Qualitatively comparisons on the high-density wheat spikes in different densities in the field, from top to bottom: 21–30, 31–40, 41–50, 51–60, 61–70, 71–80 wheat spikes per images. From left to right are input images, and the results generated by Mask R-CNN, Cascade Mask R-CNN, and Our method. The red and green box parts of each image have been enlarged in order to better show the experimental details.

concerning wheat spikes obscured by wheat leaves, our method demonstrates the capability to detect these obscured spikes. In contrast, other methods do not fully succeed in detecting the wheat spikes in such complex scenarios. On the other hand, our method also exhibits commendable performance in detecting small targets. For instance, in the 3rd column of Fig. 11. and the 3rd row of Fig. 12, our approach successfully detects small Fusarium Head Blight (FHB) lesions on the wheat spikes located at the edges of the images. In contrast, other methods fail to detect such small lesions. As demonstrated in the first row of Fig. 13, our method is also effective in detecting small lesions on wheat spikes in small-target scenarios under backlit conditions. This capability further illustrates the robustness of our approach in challenging imaging environments.

Effects of different viewing angles of photographing. To verify the generalization capability of our method, we conducted tests on wheat spike detection and segmentation in unstructured environments.

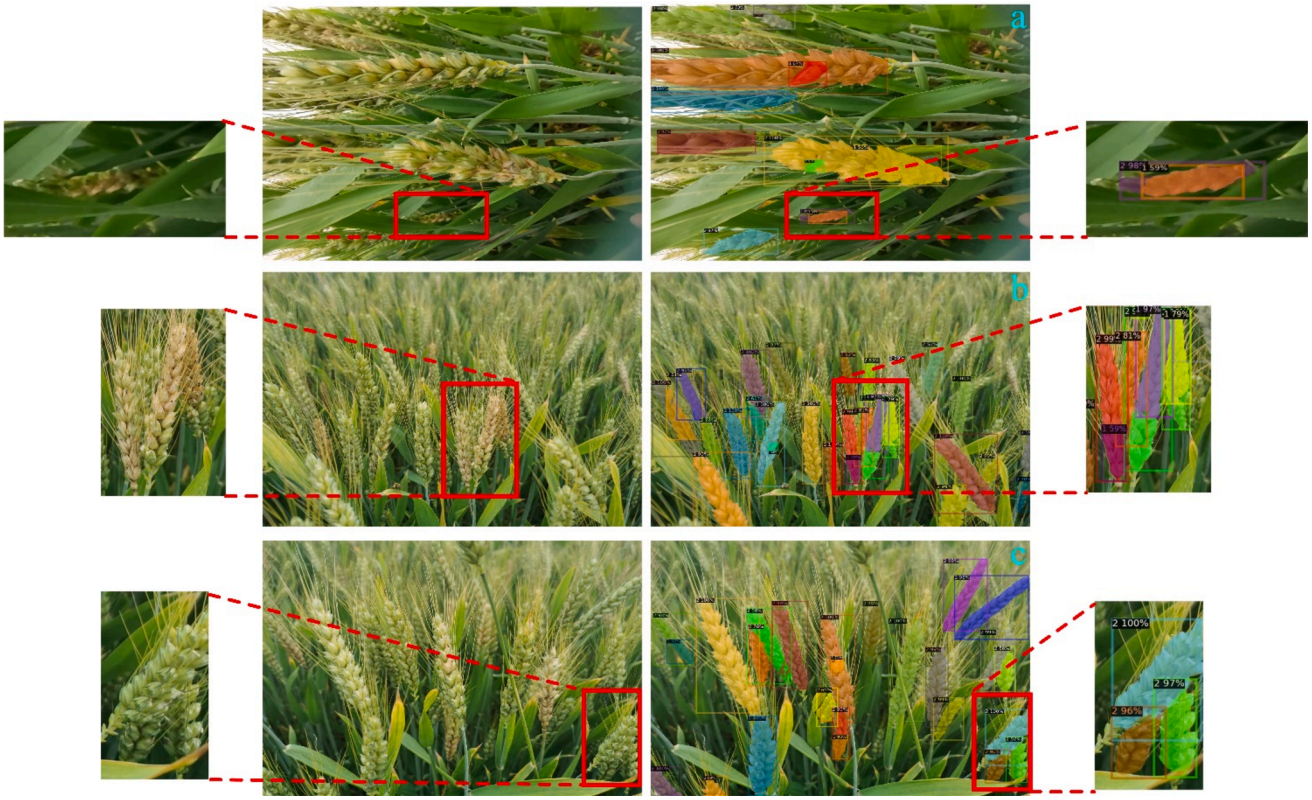

Fig. 13. Visualization results of wheat spikes and FHB segmentation under the three complex conditions. (a) Small target affected by shadows, occlusion, and uneven illuminations. (b) Spikes with FHB adjacent to each other. (c) Wheat spikes occlusion and cut at the image borders. The red box part of each image has been enlarged in order to better show the experimental details.

Visual results in scenarios with non-ideal shooting angles, target tilts, varying target scales, and overexposure are presented in Fig. 14. Our method consistently identified a greater number of wheat spikes compared to other methods, demonstrating its effectiveness in complex and varied imaging conditions.

Discussion of Limitations and Challenges One of the significant challenges in deploying deep learning for detecting Fusarium Head Blight (FHB) is maintaining high accuracy under highly variable field conditions, where factors like occlusion, variable lighting conditions, and background noise significantly complicate the detection and segmentation tasks. Our research addresses these challenges by introducing a novel deep learning architecture that integrates advanced instance segmentation techniques with a coarse-to-fine strategy. This approach not only enhances the accuracy of detecting and segmenting wheat spikes and FHB lesions in complex backgrounds but also improves the model's ability to handle specific variabilities associated with field images.

Innovatively, our model incorporates a hybrid of convolutional neural networks with transformers to leverage both local and global contextual information. This significant departure from traditional methods, which primarily rely on CNNs, allows for better feature representation and recognition accuracy, particularly in cluttered scenes. Furthermore, our method drastically reduces processing times while maintaining high throughput, which is essential for real-time agricultural decision-making.

The capability of our model to perform end-to-end segmentation and classification concurrently sets it apart from most current methodologies that focus solely on detection or segmentation. This comprehensive approach ensures that our model not only addresses the immediate needs of accurate disease detection but also facilitates broader agricultural management practices, including yield prediction and disease severity assessment.

4. Conclusions

In this paper, we introduce a high-throughput deep learning architecture for detecting and segmenting wheat spikes and Fusarium Head Blight (FHB) in complex field environments. Our approach integrates a multi-scale deep feature pyramid, quadtree-based refinement, and a transformer-based network, which significantly enhances instance segmentation accuracy, specializing in refined object segmentation and handling occlusion. This end-to-end model enables simultaneous task execution, significantly improving efficiency with an average image processing time of 3 s, ideal for large datasets and computationally intensive tasks in agricultural image analysis. Specifically, our model achieves a box AP of 64.408 and a mask AP of 64.966 on the FHB-SA dataset for wheat spike and FHB lesion segmentation. These results mark substantial improvements of 2.3 % over Mask R-CNN, 2.6 % over Cascade Mask R-CNN, and 2.1 % over PointRend. Even under more stringent IoU thresholds (AP75), our method outperforms the baselines by a margin of 3 %, maintaining superior performance. Additionally, by effectively reducing false positives caused by background interference, we have improved detection accuracy from 0.842 to 0.901. These performance metrics not only showcase our model's effectiveness but also highlight its potential in facilitating rapid and accurate diagnosis of wheat Fusarium Head Blight. The method allows for non-destructive, high-throughput, rapid, and accurate segmentation of wheat spikes and Fusarium Head Blight (FHB) lesions, saving a significant amount of human labor. Furthermore, the system assists agricultural workers in assessing diseases and predicting yields accurately, and facilitates targeted research. The approach provides technical support for automating, improving efficiency, and enhancing the accuracy of diagnosing wheat Fusarium Head Blight.

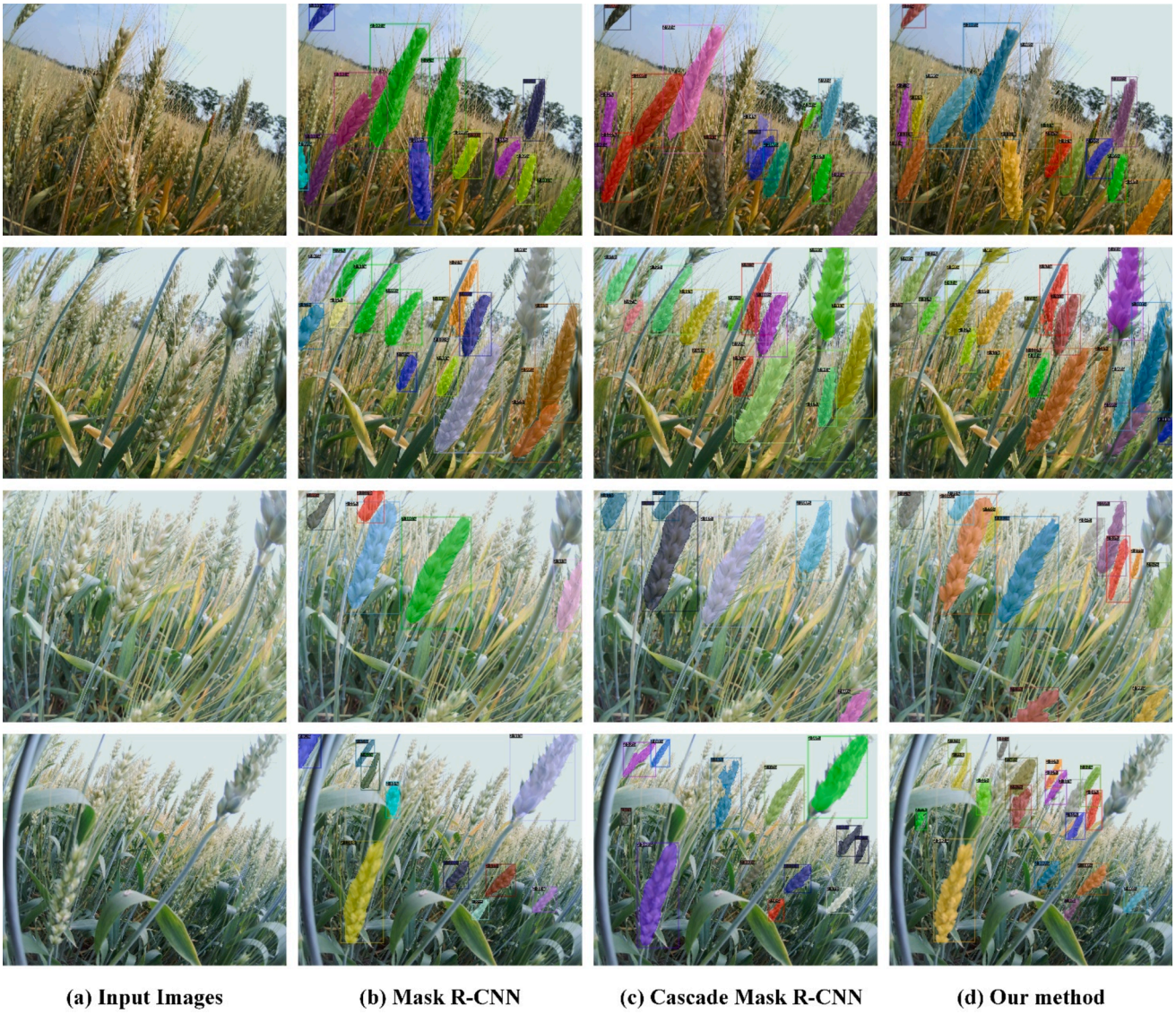

Fig. 14. Visualization results of wheat spikes and FHB segmentation under the three complex conditions.

CRediT authorship contribution statement

Qiong Zhou: Writing – original draft, Software, Methodology, Conceptualization. **Ziliang Huang:** Visualization, Validation, Software. **Liu Liu:** Writing – review & editing, Methodology. **Fenmei Wang:** Validation, Formal analysis. **Yue Teng:** Validation, Software, Investigation. **Haiyun Liu:** Visualization, Validation. **Youhua Zhang:** Writing – review & editing. **Rujing Wang:** Project administration, Funding acquisition, Data curation.

Declaration of competing interest

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

Acknowledgement

This work was supported by the National Key R&D Program of China (2023YFD1702105) and the National Key R&D Program of China (2021YFD2000205).

Data availability

The data that has been used is confidential.

References

Bao, W., Yang, X., Liang, D., Hu, G., Yang, X., 2021. Lightweight convolutional neural network model for field wheat ear disease identification. Comput. Electron. Agric. 189. https://doi.org/10.1016/j.compag.2021.106367.

Batin, M.A., Islam, M., Hasan, M.M., Azad, A., Alyami, S.A., Hossain, M.A., Miklavcic, S. J., 2023. WheatSpikeNet: an improved wheat spike segmentation model for accurate estimation from field imaging. Front. Plant Sci. 14, 1226190. https://doi.org/10.3389/fpls.2023.1226190.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 6154–6162.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., 2019. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4974–4983.

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y., 2020. Blendmask: Top-down meets bottom-up for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8573–8581.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 764–773.

Gao, C., Guo, W., Yang, C., Gong, Z., Yue, J., Fu, Y., Feng, H., 2024. A fast and lightweight detection model for wheat fusarium head blight spikes in natural environments. Comput. Electron. Agric. 216. https://doi.org/10.1016/j.compag.2023.108484.

Gao, Y., Wang, H., Li, M., Su, W.-H., 2022. Automatic tandem dual blendmask networks for severity assessment of wheat fusarium head blight. Agriculture 12. https://doi.org/10.3390/agriculture12091493.

Garg, M., Ubhi, J.S., Aggarwal, A.K., 2021. Deep learning for obstacle avoidance in autonomous driving, Autonomous driving and advanced driver-assistance systems (ADAS). CRC Press 233–246.

Gu, C., Wang, D., Zhang, H., Zhang, J., Zhang, D., Liang, D., 2021. Fusion of deep convolution and shallow features to recognize the severity of wheat fusarium head blight. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.599886.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.

Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019. Mask scoring r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 6409–6418.

Jin, X., Jie, L., Wang, S., Qi, H., Li, S., 2018. Classifying Wheat Hyperspectral Pixels of Healthy Heads and Fusarium Head Blight Disease Using a Deep Neural Network in the Wild Field. Remote Sens. (Basel) 10. https://doi.org/10.3390/rs10030395.

Ke, L., Danelljan, M., Li, X., Tai, Y.-W., Tang, C.-K., Yu, F., 2022. Mask transfiner for high-quality instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4412–4421.

Kirillov, A., Wu, Y., He, K., Girshick, R., 2020. Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 9799–9808.

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

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

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

Liu, C., Wang, K., Lu, H., Cao, Z., 2022. Dynamic color transform networks for wheat head detection. Plant Phenomics 2022. https://doi.org/10.34133/2022/9818452.

Ma, J., Li, Y., Liu, H., Du, K., Zheng, F., Wu, Y., Zhang, L., 2020. Improving segmentation accuracy for ears of winter wheat at flowering stage by semantic segmentation. Comput. Electron. Agric. 176. https://doi.org/10.1016/j.compag.2020.105662.

Qiu, R., Yang, C., Moghimi, A., Zhang, M., Steffenson, B.J., Hirsch, C.D., 2019. Detection of fusarium head blight in wheat using a deep neural network and color imaging. Remote Sens. (Basel) 11. https://doi.org/10.3390/rs11222658.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inform. Proc. Syst. 28.

Su, W.-H., Zhang, J., Yang, C., Page, R., Szinyei, T., Hirsch, C.D., Steffenson, B.J., 2020. Automatic evaluation of wheat resistance to fusarium head blight using dual mask-RCNN deep learning frameworks in computer vision. Remote Sens. (Basel) 13. https://doi.org/10.3390/rs13010026.

Sun, J., Yang, K., Chen, C., Shen, J., Yang, Y., Wu, X., Norton, T., 2022. Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network. Comput. Electron. Agric. 193. https://doi.org/10.1016/j.compag.2022.106705.

Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, pp. 9627–9636.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.-M., 2021. Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition. IEEE, pp. 13029–13038.

Wu, Y., Kirillov, A., Massa, F., et al., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Y., He, K., 2018. Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). IEEE, pp. 3–19.

Yang, B., Gao, Z., Gao, Y., Zhu, Y., 2021. Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module. Agronomy 11, 1202.

Zhang, D., Gu, C., Wang, Z., Zhou, X., Li, W., 2021. Evaluating the efficacy of fungicides for wheat scab control by combined image processing technologies. Biosyst. Eng. 211, 230–246. https://doi.org/10.1016/j.biosystemseng.2021.09.008.

Zhang, D.-Y., Luo, H.-S., Wang, D.-Y., Zhou, X.-G., Li, W.-F., Gu, C.-Y., Zhang, G., He, F.-M., 2022a. Assessment of the levels of damage caused by Fusarium head blight in wheat using an improved YoloV5 method. Comput. Electron. Agric. 198. https://doi.org/10.1016/j.compag.2022.107086.

Zhang, D.-Y., Luo, H.-S., Cheng, T., Li, W.-F., Zhou, X.-G., Wei, G., Gu, C.-Y., Diao, Z., 2023. Enhancing wheat Fusarium head blight detection using rotation Yolo wheat detection network and simple spatial attention network. Comput. Electron. Agric. 211. https://doi.org/10.1016/j.compag.2023.107968.

Zhang, J., Min, A., Steffenson, B.J., Su, W.H., Hirsch, C.D., Anderson, J., Wei, J., Ma, Q., Yang, C., 2022b. Wheat-net: an automatic dense wheat spike segmentation method based on an optimized hybrid task cascade model. Front. Plant Sci. 13, 834938. https://doi.org/10.3389/fpls.2022.834938.

Zhang, D., Wang, D., Gu, C., Jin, N., Zhao, H., Chen, G., Liang, H., Liang, D., 2019. Using neural network to identify the severity of wheat fusarium head blight in the field environment. Remote Sens. (Basel) 11. https://doi.org/10.3390/rs11202375.

Zhao, J., Cai, Y., Wang, S., Yan, J., Qiu, X., Yao, X., Tian, Y., Zhu, Y., Cao, W., Zhang, X., 2023. Small and oriented wheat spike detection at the filling and maturity stages based on wheatnet. Plant Phenomics 5, 0109. https://doi.org/10.34133/plantphenomics.0109.

Zhou, Q., Huang, Z., Zheng, S., Jiao, L., Wang, L., Wang, R., 2022. A wheat spike detection method based on transformer. Front. Plant Sci. 13, 1023924. https://doi.org/10.3389/fpls.2022.1023924.