Wheat maturity identification based on improved RT-DETR model

Mingyue Yan[1], Jingfa Yao[2,3]*, Guifa Teng[1,4,5]*

[1]School of Information Science and Technology, Hebei Agricultural University, Hebei, China; [2]Hebei Software Engineering Department, Baoding, China; [3]Hebei College of Intelligent Interconnection Equipment and Multi-Modal Big Data Application Technology Research and Development Center, Baoding, China; [4]Hebei Digital Agriculture Industry Technology Research Institute, Hebei, China; [5]Key Laboratory of Agricultural Big Data in Hebei Province, Hebei, China

***Corresponding Authors:** Jingfa Yao, Hebei College of Intelligent Interconnection Equipment and Multi-Modal Big Data Application Technology Research and Development Center, Baoding, China. Email: yaojingfa@hbsi.edu.cn and Guifa Teng, School of Information Science and Technology, Hebei Agricultural University, Hebei, China. Email: tguifa@hebau.edu.cn

Academic Editor: Ravi Pandiselvam, PhD, Scientist, ICAR-Central Plantation Crops Research Institute, Kerala, India

Received: 16 October 2024; Accepted: 17 June 2025; Published: 17 November 2025

ORIGINAL ARTICLE

Abstract

As a key global cereal crop, wheat undergoes distinct growth phases and developmental stages that are critical for informing agricultural management strategies. Accurate assessment of wheat maturity, a crucial determinant of yield potential, is vital for optimizing harvest timing, ensuring consistent grain quality, and maximizing economic returns in production systems. Traditional maturity evaluation methods, which primarily rely on manual field inspections and subjective visual scoring, have inherent limitations, such as these being labor-intensive and time-consuming, requiring expert knowledge, exhibiting inconsistent inter-observer reproducibility, and lacking scalability for large-scale monitoring. Recent advances in artificial intelligence (AI) have revolutionized agricultural monitoring by addressing these constraints. Deep learning-based computer vision techniques now enable automated maturity assessment frameworks, utilizing algorithmic pattern recognition of spectral and morphological features in crop imagery. These systems outperform human experts in terms of efficiency—processing thousands of images per hour—and accuracy, with error margins of less than 2% in controlled trials. Moreover, these AI-driven systems facilitate data-driven decision-making for precision irrigation, nutrient management, and yield forecasting. Their integration with unmanned aerial vehicles and internet of things-enabled edge devices allows for real-time, field-deployable solutions that minimize operational costs and resource inputs. This technological convergence is fostering the development of precision agriculture ecosystems, where AI-based maturity analytics drive sustainable intensification practices, ranging from genotype selection to post-harvest logistics, thereby advancing global food security initiatives.

Keywords: maturity; wheat ear recognition; residual network; attention mechanism; RT-DETR

Introduction

Wheat, a globally significant cereal crop, undergoes growth processes and developmental stages that are pivotal for agricultural management and decision-making. The maturation of wheat directly impacts its yield; therefore, an accurate assessment of wheat maturity is essential for farmers. This not only optimizes yield but also ensures quality and stability, thereby enhancing the economic returns of agricultural produce. Traditional methods of evaluating wheat maturity have primarily relied on manual observations and field surveys.

 ISSN 1757-837X online, DOI 10.15586/qas.v17i4.1565

However, these approaches are constrained by low efficiency, high costs, and susceptibility to subjective biases, failing to capture critical maturity data directly. With the advent and widespread integration of artificial intelligence (AI) technologies, deep learning-based methods for assessing wheat maturity are increasingly replacing traditional manual techniques. Deep learning improves detection efficiency and precision through algorithmic learning and image analysis, enabling large-scale monitoring and management, reducing costs, and increasing benefits. This technological transition is driving the evolution of smart agriculture, ushering in a wave of technological innovations and new opportunities for agricultural production.

In recent years, the rapid advancement of deep learning technologies has led to significant progress in the field of object detection. Both domestic and international researchers have made crucial contributions to the detection of wheat maturity. Meng conducted a study in Yucheng City, Shandong, China, where specific observation points were selected, and a remote sensing prediction model for the maturity period of winter wheat was developed using HJ-1A CCD and HJ-1B IRS data (Meng *et al.*, 2011). This model successfully enabled the remote sensing detection and prediction of the maturity period of winter wheat. However, the data collection and regression model development stages of this method are time-consuming (Meng *et al.*, 2011). Du (2019) employed digital image data to monitor quantitatively the wheat filling process and maturity by constructing a reverse model for post-flowering days and a maturity index at the end of grain filling, utilizing superpixel block scale based on spike color characteristics combined with color and temperature features. Su (2011) explored the underlying theories and methodologies of machine vision and hyperspectral technology for grain and impurity recognition, successfully predicting the optimal harvest time for wheat. Yang *et al.* (2021) proposed the use of unmanned aerial systems (UAS) to capture aerial images, accurately predicting the wheat maturity period with an accuracy rate of 88.6%. Darvishzadeh suggested leveraging high-resolution drone data to monitor small-scale crops at sub-centimeter spatial resolution, achieving a maturity detection accuracy of 87% (Li *et al.*, 2022). These studies underscore the widespread application and continuous optimization of machine learning (ML) in wheat spike detection, signaling its potential to provide more effective tools and technical support for agricultural production, thereby significantly enhancing the accuracy and efficiency of yield predictions.

Accurate assessment of crop growth status is crucial for improving both quality and yield of agricultural products. With the rapid development of deep learning technologies, object detection has experienced significant breakthroughs in the industrial sector. Specifically, in the domain of wheat spike detection, object detection models have made remarkable progress and have become central to wheat cycle classification and yield estimation. These models now lead the industry in terms of both detection accuracy and speed (Jin *et al.*, 2022; Kumar and Kukreja, 2022; Xu *et al.*, 2023). Precise prediction of wheat maturity is essential for the efficient utilization of resources and optimization of yield and quality. Through the application of accurate maturity prediction technologies, wheat can be harvested at its peak nutritional content and maturity, maximizing grain weight and substantially enhancing the overall quality. Furthermore, by preventing over-ripening, key measures can effectively reduce natural field shattering, directly increasing the harvestable yield.

Timely harvest predictions not only minimize field losses but also optimize harvesting schedules, facilitating a more efficient use of agricultural machinery and labor. This, in turn, alleviates resource constraints and reduces costs during peak harvest periods. From an economic perspective, the formulation of effective harvesting strategies can significantly decrease unnecessary expenses in agricultural production. This integrated management approach not only improves the economic sustainability of wheat production but also underscores the progress in agricultural technology, highlighting the critical role of precision agricultural technologies in modern agriculture.

In the field of wheat maturity detection, several challenges remain to be addressed. For example, during the wax ripening stage, the color of wheat leaves is similar to that of the wheat ears, which can lead to misdetections and omissions. Additionally, phenomena such as ear occlusion and overlapping further complicate the detection of wheat maturity. These issues necessitate the development of more efficient and precise technological solutions to optimize the entire wheat maturity detection process. The Real-Time DEtection TRansformer (RT-DETR) utilizes a transformer architecture to facilitate global feature interaction, overcoming the limitations of manually designed anchor boxes in convolutional neural networks (CNNs). RDE-DETR, the enhanced RT-DETR model, is particularly well suited for small object detection, especially in dense wheat ear scenarios. Its hybrid encoder architecture, anchor-free design, multi-scale feature fusion, and optimized training strategies systematically address the challenges of feature ambiguity, localization bias, and computational efficiency in small object detection tasks.

Data and Methods

Data collection and dataset construction

The experimental dataset for this study was obtained from Jingxiu District, Baoding City, Hebei Province, China, utilizing the wheat cultivar Shannong 28. Data collection occurred between 5 May 2023 and 15 June 2023, encompassing critical growth stages of wheat, from heading to wax ripening. The primary tools used for data capture included selfie sticks and iPhone 13 smartphones, with data being collected daily between 0800 am and 06:00 pm. In total, 2,000 spike images were captured, all stored in JPEG format. The image acquisition process involved the operator standing on the ground, utilizing selfie sticks and smartphones for capturing images. To ensure the dataset's diversity and comprehensiveness, multiple shooting conditions were incorporated, including variations in lighting angles, introduction of image noise, and image classification under different occlusion scenarios. These efforts aimed to replicate the diverse conditions encountered in real-world field environments (Table 1). Detailed methodology is outlined below:

(1) *Classification under different lighting conditions*: Based on the time of image-capture, it was observed that the lighting at 8:00 am and 6:00 pm is relatively weak whereas between 12:00 noon and 3:00 pm, sunlight intensity is significantly higher. To more accurately evaluate the effect of lighting on image quality, the images were classified into two categories: (a) "strong light conditions" and (b) "shadow conditions." Specific examples, as depicted in Figures 1A–1F, illustrate the appearance of wheat spikes under different lighting scenarios.

(2) *Image enhancement*: In order to enhance diversity of the dataset and test the model's robustness under various conditions, noise processing was applied to the captured images of wheat spikes. Specifically, Gaussian noise was added to a portion of images to simulate sensor noise or uneven lighting effects that may occur in real

Table 1. Data shooting information.

Influencing factors		Image category	Number of images/pieces
Light intensity	Heading stage	Direct sunlight	260
		Against sunlight	251
	Grain filling stage	Direct sunlight	243
		Against sunlight	236
	Maturity stage	Direct sunlight	242
		Against sunlight	269
Smoothness level		Smooth	1,670
		Blur and noise	700

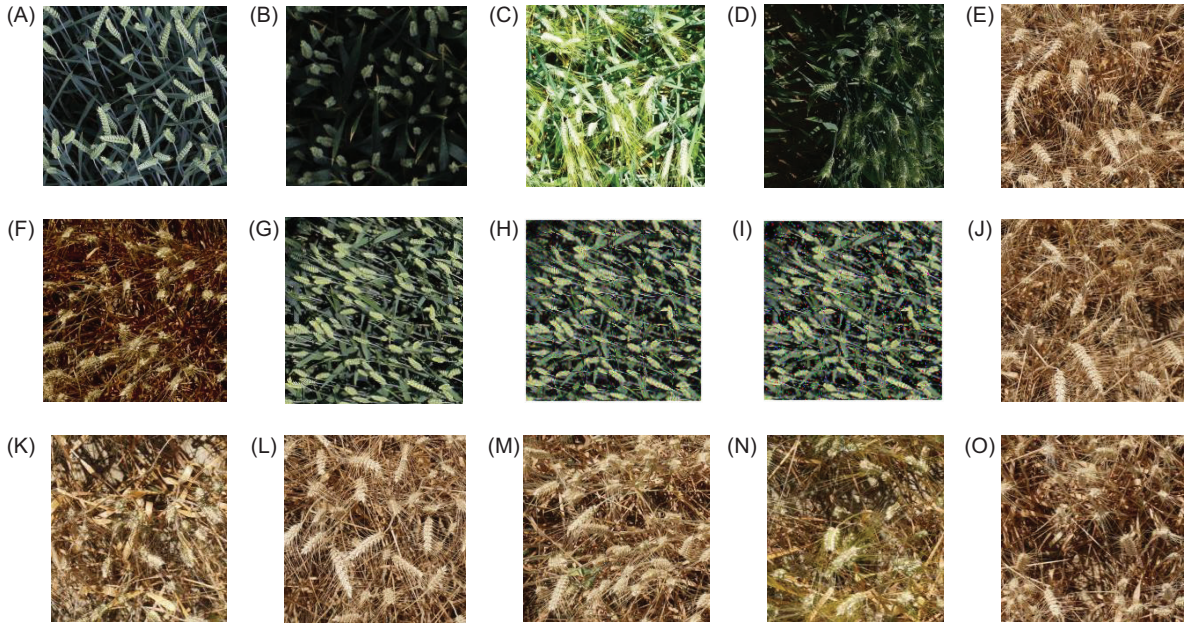

Figure 1. Dataset classification. (A) Heading stage with direct sunlight. (B) Heading stage against sunlight. (C) Grain filling stage with direct sunlight. (D) Grain filling stage against sunlight. (E) Maturity stage with direct sunlight. (F) Maturity stage against sunlight. (G) Translates to "original image." (H) Translates to "Gaussian noise." (I) Salt-and-pepper noise. (J) Shandong Province "Zhongxin Mai 999." (K) Shandong Province "Shannong 28." (L) Shandong Province "Tobacco Farmer 999" (M) Hebei Province "Zhongxin Mai 999." (N) Hebei Province "Shannong 28." (O) Hebei Province "Tobacco Farmer 999."

environments. Additionally, salt-and-pepper noise was applied to another subset of images to simulate random disturbances that might occur during image capture, such as dust or water droplets.

In all, 2,370 wheat spike images were acquired, following the meticulous classification and organization. These images were subsequently annotated using the LabelImg software. During the annotation process, wheat spikes were identified, and labels corresponding to their maturity stages were assigned: "m" for mature, "h" for heading, and "g" for grain filling. Each annotated image generated an associated XML file containing the label information, facilitating subsequent model training and evaluation. For systematic training and assessment of the improved RT-DETR model, the annotated dataset was partitioned into training, validation, and test sets at an 8:1:1 ratio. The training set comprised 1,890 images, while both validation and test sets contained 240 images each.

Comparison of datasets

In order to verify the generalization of the RT-DETR model, experiments were conducted on public datasets, self-made datasets, and mixed datasets of public and self-made datasets. The experimental results are shown in Table 2. The self-made dataset had the highest accuracy, reaching 97.3%, followed by the public dataset, and the mixed dataset had the lowest accuracy.

Methodology research

RT-DETR network models

In the field of object detection, while deep learning technologies have made significant progress, current object detection algorithms still face substantial challenges when dealing with complex scenarios, such as occlusions, lighting variations, and changes in object scale (Ma *et al.*, 2020; Zhong *et al.*, 2019). Moreover, the robustness and generalization capabilities of these algorithms must be further enhanced to ensure effective performance across diverse environments and tasks. To address these issues, it is imperative for researchers to refine algorithm architectures, improve adaptability to environmental changes, and strengthen their applicability in various agricultural contexts. These improvements not only increase the practical efficacy of the models but also contribute to the advancement of smart agriculture technologies (Wang *et al.*, 2020).

Table 2. Comparison of datasets.

Data set	P	R	F1	mAp
Common	96.9	96.4	96.5	97.8
Self-made	97.3	97.0	97.4	98.3
Mix	96.1	96.5	96.2	97.3

Notes: P: precision; R: recall; mAp: mean average precision.

In response to these challenges, various innovative algorithms are proposed, including the YOLO series, RT-DETR, and EfficientDet. The YOLO series, as a classic object detection framework, is widely recognized for its balanced performance in terms of detection speed, accuracy, and small object detection (Khaki *et al.*, 2022; Zhang *et al.*, 2022). However, in the context of complex backgrounds, such as wheat fields, YOLO's performance may degrade when detecting small objects such as wheat spikes. The EfficientDet model, built on the EfficientNet backbone, utilizes a compound scaling strategy that enhances both efficiency and accuracy in object detection tasks. Its adaptive features allow for robust performance across devices with varying computational capabilities (Jia *et al.*, 2022). Nonetheless, the EfficientDet model may require additional fine-tuning to achieve optimal results in specific complex scenarios. In contrast, the RT-DETR model leverages the self-attention mechanism of transformers to effectively process contextual information in images, eliminating the need for traditional anchor boxes and directly outputting target positions and class labels (Liu *et al.*, 2024; Zhu and Kong, 2024). This design improves the model's simplicity and robustness. However, in situations involving multiple overlapping objects or partially occluded targets, the model's performance may be compromised, as the attention mechanism might not fully capture the information of occluded objects.

To address these challenges, the present study focuses on optimizing the model from two key aspects: first, enhancing the model's feature extraction capabilities to ensure accurate identification and localization of targets of varying sizes in complex environments; second, implementing model lightweighting to reduce computational demands, enabling broader applicability across different scenarios and devices. These strategies aim to improve the model's generalization and robustness, ensuring its effectiveness and reliability in practical agricultural applications.

Enhancement and deployment of RT-DETR network model

The enhanced RT-DETR model is primarily composed of three key components: the backbone network, the neck encoding network, and the decoding prediction network. The architecture of this improved model is depicted in Figure 2. This model replaces the traditional backbone network with a ResNet18 residual network, reducing the model's memory footprint. Additionally, the convolution

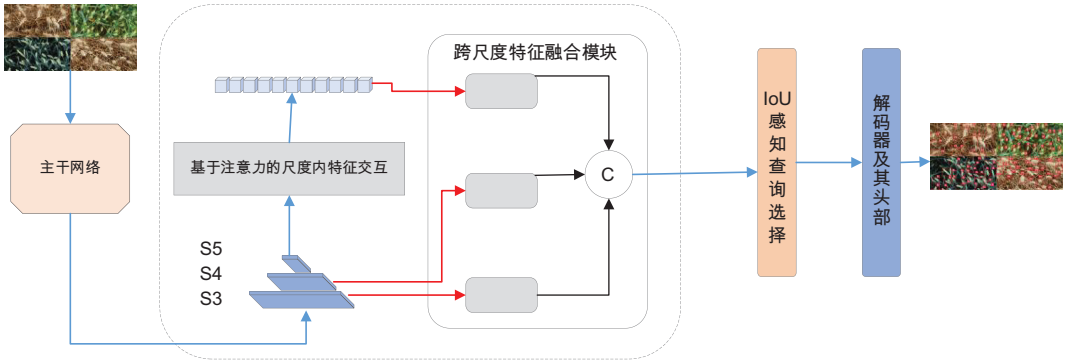

Figure 2. Schematic of RDE-DETR network structure.

operations within the ResNet18 network are replaced by Dynamic Separable Convolution (DSConv), providing sufficient depth for capturing intricate image features while preserving computational efficiency. To mitigate the excessive computational overhead introduced by the extended window multi-head self-attention operation in the ResNet18 residual network, the cascaded grouped attention mechanism from EfficientViT is integrated. This modification enhances the multi-head self-attention mechanism within the inverse residual moving module, boosting both information capacity and feature diversity while simultaneously reducing computational redundancy.

The RDE-DETR model first scales and pads the input image to a predefined resolution before feeding it into the network for inference preparation. The backbone network, utilizing a convolutional neural network, extracts key features from the image and produces outputs at three high-level stages (S3, S4, and S5). These outputs are then passed to the neck encoding network, which transforms the multi-level features into a sequence of image feature vectors by interacting and fusing features at various scales. The decoding prediction network employs a query selection mechanism based on intersection over union (IoU), extracting a fixed number of feature sequences from the neck encoding network's results as the initial object queries for the decoder. The decoder, equipped with an auxiliary prediction head, iteratively refines these object queries, ultimately generating precise prediction boxes. The challenges of feature ambiguity, localization errors, and computational inefficiency in small object detection are effectively tackled through the adoption of a hybrid encoder architecture, anchor-free design, multi-scale feature fusion, and optimized training methodologies.

Unmanned aerial vehicles (UAVs) or field-mounted towers, integrated with high-definition or multispectral cameras, are capable of conducting scheduled and targeted image acquisition within wheat fields, capturing high-resolution images of wheat ears. These images are transmitted either in real-time or in batches to a cloud server for processing through wireless communication technologies, such as 5G, Wi-Fi, or LoRa. On the server end, the transmitted images undergo processing via an enhanced RT-DETR visual object detection model, which precisely identifies the regions containing wheat ears and extract-critical attributes, including shape, position, and other relevant characteristics of each ear. The features extracted from the RT-DETR model, such as color, texture, and shape, serve as input data for the subsequent maturity analysis model. This model, trained on historical data, establishes a mapping between wheat maturity and its associated image features, enabling precise maturity assessment of each wheat ear across various growth stages. This methodology offers an automated, high-precision solution for wheat growth monitoring, thus augmenting the intelligence and decision-making capabilities within agricultural management (Figure 3).

To replace ResNet18 backbone in RT-DETR model

The selection of backbone network is pivotal to the overall performance of object detection models, making crucial the choice of appropriate network architecture. As a result, the Residual Network (ResNet) was adopted to replace the original network structure. ResNet18, a lightweight variant within the ResNet family, is noted for its simplicity and efficiency, consisting of 18 weight layers. This model is a preferred choice for improving detection performance (Helaly *et al.*, 2023; Hu *et al.*, 2024). To assess the performance of various backbone networks, this study compares VanillaNet, developed by the Huawei team, ResNet, the classic VGG network, and HGNetV2, which is used in the original RT-DETR model. Evaluation metrics include model parameter count and mean precision. The results reveal that ResNet18 has only 19.2 M parameters, significantly fewer than the other networks, while maintaining a high average precision. These findings not only validate the effectiveness and robustness

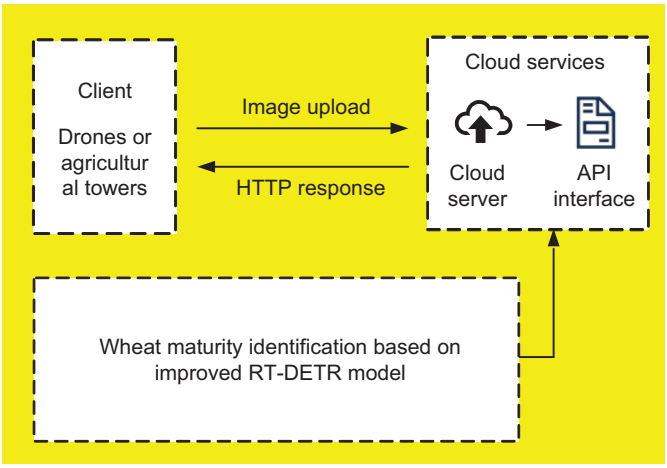

Figure 3. Model deployment.

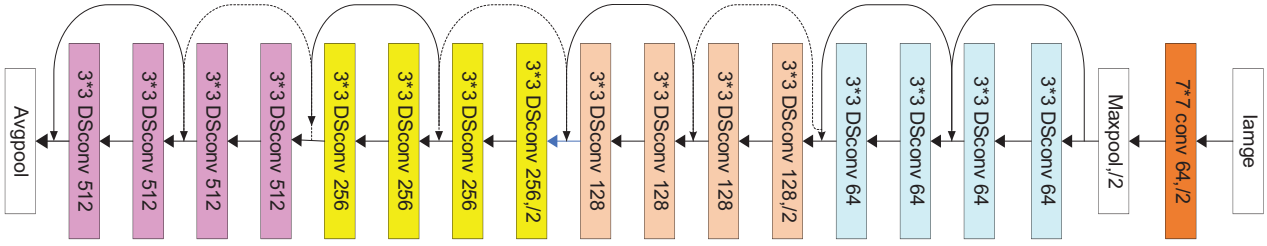

Figure 4. ResNet18 improved network architecture diagram.

of selecting ResNet18 as the backbone network but also emphasize its advantage of a reduced parameter count. This reduction simplifies the computational structure, decreases storage requirements, and lowers computational resource consumption, making it particularly well suited for practical applications.

To replace convolutional layers in ResNet18

Although ResNet18 is favored for its high accuracy and low parameter count, its reliance on convolution operations for feature extraction exhibits some limitations when addressing specific issues. The convolution kernels of ResNet18 focus on extracting features from local neighborhoods, which restrict their ability to capture long-range dependencies. For applications requiring the detection of fine changes in color and shape, such as wheat spike detection, this local perception may be insufficient. Moreover, owing to the fixed receptive field of traditional convolutions, their capability to detect small targets or handle multi-scale occlusions is also limited.

To address these challenges, this paper introduces a novel convolution technique, Dynamic Separable Convolution (DSConv). DSConv is an innovative approach that enhances the flexibility of convolutional kernels by incorporating deformable offsets, thereby enabling more precise capture of complex geometric features in images while reducing computational overhead and accelerating inference speed (Yin *et al.*, 2023). The primary advantage of this method lies in its ability to adjust dynamically the receptive field to accommodate varying feature scales in the image, making it especially effective for detecting slender or irregularly shaped objects. Figure 4 presents a schematic of the ResNet18 network following convolution replacement. Table 3 demonstrates a significant improvement in accuracy and a notable reduction in computational complexity after replacing the convolution operation.

Table 3. Comparison of datasets.

Convolution type	mAP(%)	FLOPs(G)	Testing time (ms)
Standard convolution	93.6	12.5	30.2
DSConv	97.1	9.1	29.7

Note: DSConv: dynamic separable convolution; mAp: mean average precision.

DSConv introduces deformable offsets, allowing the convolutional kernels in standard 2-dimensional (2D)

convolution operations to be more adaptable. This design is inspired by previous studies, but DSConv makes key improvements in its implementation, making it more suitable for handling complex scenes in the real world (Chen *et al.*, 2024; Yao *et al.*, 2024). Specifically, DSConv not only changes the spatial layout of the convolutional kernels but also ensures through an iterative strategy that the dynamic adjustment of the receptive field closely corresponds to the target features, overcoming the limitations of traditional convolutions in dealing with images with complex backgrounds. Additionally, DSConv maintains continuous attention on the target through a gradual refinement process, which is particularly important for achieving precise detection. Through this strategy, DSConv effectively resolves the issues of occlusion and detection of small targets encountered in specific application scenarios, thereby enhancing the overall performance and applicability of the model (Liu et al., 2023).

Figure 5 illustrates the coordinate architecture and receptive field of DSConv. Initially, a standard 2D convolution is represented using coordinates denoted by K, with the central coordinate denoted by $K_i = (x_i, y_i)$. For a 3×3 convolution kernel K and a dilation rate set to 1, it can be represented as follows:

$$K = \{(x-1, y-1), (x-1, y), ..., (x+1, y+1)\} \quad (1)$$

In order to provide greater flexibility to convolution kernels in focusing on complex geometric features of targets, deformable offsets D are introduced. However, allowing the model to learn freely these offsets might lead to the receptive field deviating from the target. To address this issue, DSConv adjusts convolution kernels along the *x*-axis and *y*-axis,

$$K_{i\pm c} = \left\{ \begin{array}{l} (x_{i+x}, y_{i+c}) = \left(x_i + c, \sum_i^{i+c} \Delta y \right) \\ (x_{i-c}, y_{i-c}) = \left(x_i - c, \sum_{i-c}^{i} \Delta y \right) \end{array} \right\}. \quad (2)$$

The coordinates along the *y*-axis are as follows:

$$K_{j\pm c} = \left\{ \begin{array}{l} (x_{j+x}, y_{j+c}) = \left(x_j + \sum_j^{j+c} \Delta x, y_{j+c} \right) \\ (x_{j-c}, y_{j-c}) = \left(x_j - \sum_{j-c}^{j} \Delta y, y_{j-c} \right) \end{array} \right\}, \quad (3)$$

where Σ represents the cumulative deformation. Assuming Δ is typically fractional, bilinear interpolation is used to handle this, as shown in Equation (4),

$$K = \Sigma_K B(K', K) \cdot K'. \quad (4)$$

In this context, *B* represents the bilinear interpolation kernel, which can be decomposed into two separate linear interpolation kernels. Here, *b* denotes a single linear interpolation kernel as indicated in Equation (5),

$$B(K, K') = b(K_x, K_x') \cdot b(K_y, K_y') \quad (5)$$

Standard convolution is performed using a kernel size of 3×3×C, while depthwise separable convolution consists of two kernels, sized 3×3×1 and 1×1×C, as shown in Figure 3. The ratio of depthwise separable convolution to standard convolution is presented in Equation (6). Compared to standard convolution, depthwise separable convolution has fewer parameters. In Equation (6), *R* represents the ratio of the number of parameters, *N* denotes the number of input channels, and *k* symbolizes the size of the convolutional kernel,

$$R = \frac{1}{N} + \frac{1}{k^2}. \quad (6)$$

Introduction and visualization of attention mechanism

In the RT-DETR model, the uniform weighting approach applied to all features can lead to missed and false detections of small-scale targets, such as wheat ears. To address this issue, this study introduces an improved

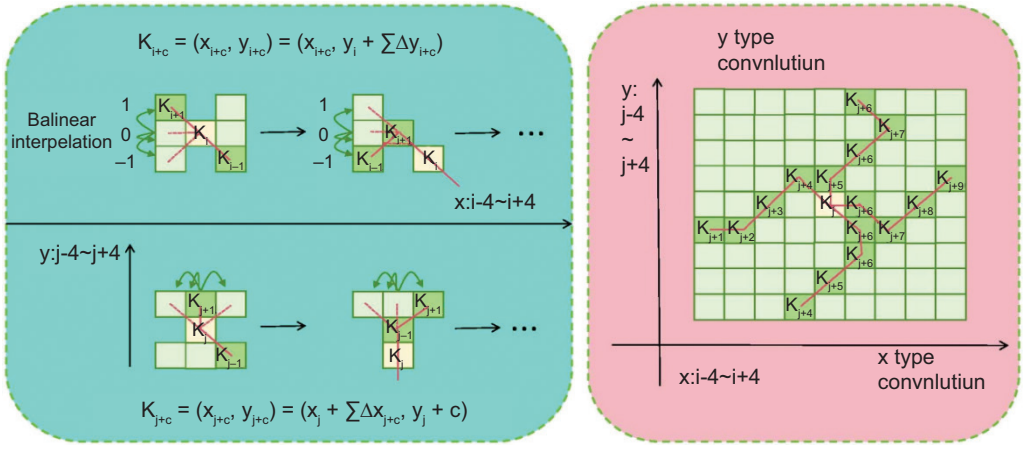

Figure 5. Left: Illustration of coordinates calculation of DSConv. Right: The receptive field of DSConv.

multi-head self-attention mechanism (MHSA), which effectively enhances the weights of occluded or small targets in the feature map through a channel attention module, thereby making this critical information more readily identifiable and learnable by the network. However, the traditional multi-head self-attention involves extensive computations during the expansion window operations, which may lead to significant memory consumption. To this end, we further incorporate a cascaded grouped attention mechanism from EfficientViT into MHSA to optimize computational efficiency and reduce memory usage. The specific structure of this cascaded grouped attention mechanism is depicted in Figure 6.

In the improved multi-head self-attention mechanism, before computing the queries (Q), keys (K), and values (V), the attention heads are first divided into multiple groups, where Q represents the query features, K represents the key features, and V represents the value features (Zang *et al.*, 2022; Zhong *et al.*, 2023). This division allows each attention group to process independently a portion of information, reducing the data load and complexity handled by a single attention head. The results from each grouped attention head are not isolated but are cascaded, meaning the output of each sub-head serves as the input for the next, ensuring effective information transfer across sub-heads and enhancing the model's expressive capacity.

The specific operation of attention head slicing can be represented by Equation (7), while the cascaded computation process is described by Equation (8). This design not only improves processing efficiency but also, by handling information at a finer granularity, enhances the model's capability to detect targets in small scales and complex backgrounds. Through the integration and application of these techniques, the model's detection accuracy and robustness for critical small targets in the agricultural field, such as wheat ears, are enhanced effectively.

$$\begin{aligned} \tilde{X}_{ij} &= \text{Attention}(X_{ij}M_{ij}^{Q}, X_{ij}M_{ij}^{K}, X_{ij}M_{ij}^{V}) \\ \tilde{X}_{i+1} &= \text{Concat}[\tilde{X}_{ij}]_{j=1:h}M_{i}^{p}. \end{aligned} \tag{7}$$

$$X'_{ij} = X_{ij} + \tilde{X}_{i(j-1)}. \tag{8}$$

In the equation, Output X_{ij} represents the output features of the i-th input feature processed by the j-th attention head ($1 < j \leq h$), and Attention denotes the attention mechanism operation. Slice X_{ij} refers to the j-th slice of the input feature. The terms M_{ij}^{Q}, M_{ij}^{K}, and M_{ij}^{V} refer to the query features, value features, and related features obtained from mapping the input features across different layers, respectively. The term $\tilde{X}_{i+1}$ is the result of concatenating all features processed by the attention heads and then projecting them. Concat is the concatenation operation, h is the total number of attention heads, M_{i}^{p} is the projection of the features after concatenation, and X'_{ij} represents the summation of the current attention head's input slice X_{ij} and the output from the previous head $\tilde{X}_{i(j-1)}$. The cascaded grouped attention mechanism, by employing a split-and-concatenate approach, allows each attention head to capture local information while also accessing complete information from the preceding subheads. This not only enhances the capacity of the information but also maintains the richness of the features, thereby reducing computational redundancy (Li *et al.*, 2021; Zhu et al., 2022).

The application of the Cascaded Group Attention Mechanism in wheat ear detection tasks effectively captures key information in images and improves the model's recognition accuracy of wheat ears. By visualizing the

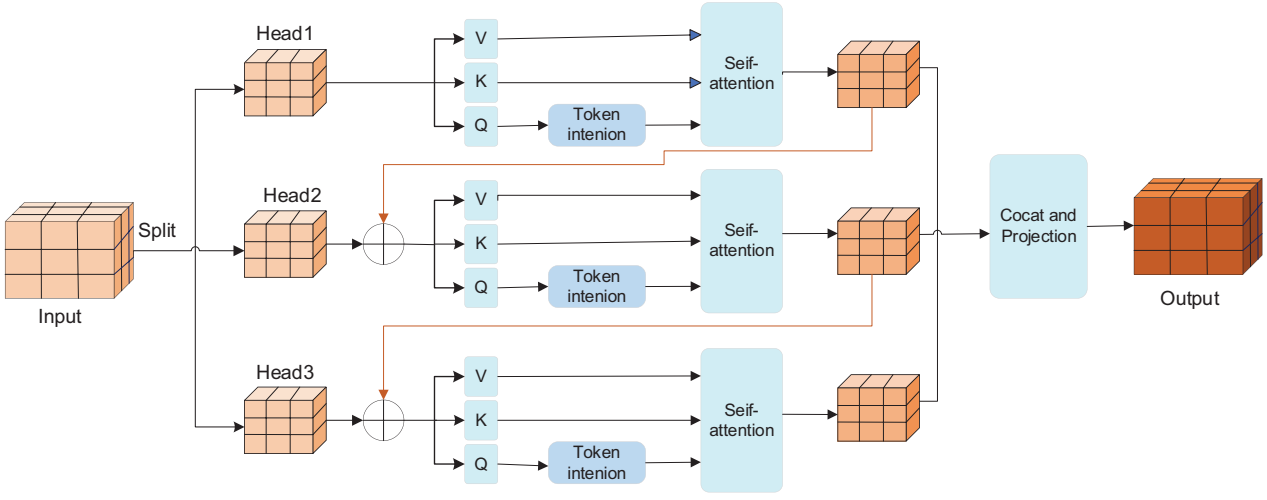

Figure 6. Cascaded group attention structure.

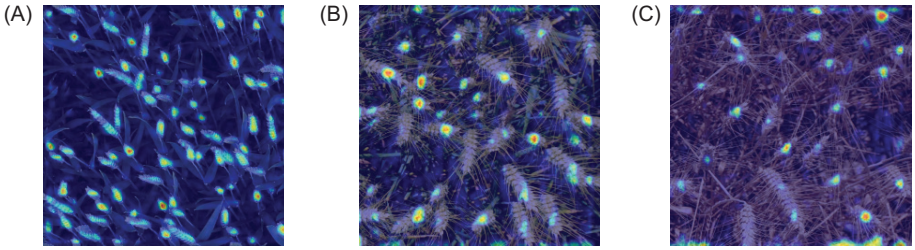

Figure 7. (A) Heat map of heading stage. (B) Thermal map during grouting period. (C) Mature heatmap.

attention heatmap, we can gain a deeper understanding of how the model focuses on different layers and regions of wheat ears. The heatmap typically uses a color gradient from cool to warm tones to represent variation in attention weights. Higher attention weights (usually represented by red or yellow) indicate areas where the model focuses more, while lower weights (typically shown in blue or green) suggest that the model pays less attention to those regions. In wheat ear images, the heatmap reveals the model's varying attention intensity across different spatial locations. During detection, different parts of the wheat ear attract different levels of attention. Especially in cases where wheat ears are densely packed or partially occluded, the model dynamically adjusts its focus through the attention mechanism, ensuring that each part of the wheat ear is effectively detected. Figure 7 presents the detailed heatmap.

Experiments and Analysis of Results

Experimental environment

The experimental environment is set up in Autodl workstation, and the main hardware configuration is shown in Table 4.

Through experimental validation, the optimization of experimental parameters has been achieved synergistically: an initial learning rate of 0.01 ensures gradient stability during AdamW optimization while maintaining alignment with transformer-based detection benchmarks. The momentum coefficient of 0.937, inherited from the YOLO optimization protocol, demonstrates superior noise suppression capabilities during small batch training. The 640×640 input resolution effectively balances GPU memory efficiency with the preservation of smaller targets, while the batch size of 16 optimizes batch normalization performance within the 24-GB VRAM constraint. The linear scaling rule, along with its implicit regularization benefits, has been adhered to. Extended training over 800 epochs addresses the long-term convergence demands inherent to transformer architectures.

Table 4. Experimental environment.

Configure	Parameters
CPU	Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60 GHz
Random access memory (RAM)	80 GB
GPUs	GeForce RTX 3090
Display memory	24 GB
Training environment	CUDA 11.8
Operating system	Ubuntu 20.04
Development environment (computer)	Python 3.8.10 Pytorch 2.0.0

Table 5. Ablation experiment.

Models	P	R	F1	mAp	Memory footprint (MB)
RT-DETR	94.9	95.5	96.5	96.8	61.3
RD-DETR	96.1	96.9	97.2	97.1	30.1
RDE-DETR	97.3	96.3	96.1	97.2	33.7

Notes: P: precision; R: recall; mAp: mean average precision.

Analysis of results

Ablation experiment

In order to better verify improvement of the network model by improvement points, the original model RT-DETR, the replacement of the backbone network and the replacement of the convolution model RD-DETR, and the model RDE-DETR with the addition of a cascaded group attention mechanism were used. The specific experimental results are shown in Table 5. It is observed that after replacing network backbone and convolution, the model performance improved significantly. While ensuring a reduction in memory usage, the P and mAp values increased by 1.2 and 0.3 percentage points, respectively. By introducing a cascaded group attention mechanism, although it increases the memory footprint slightly, the overall performance is improved while reducing computational redundancy.

Comparison results of different detection models

In order to ensure fairness in the experiment, the same dataset was used across all trials. The training process spanned 400 epochs, with initial fluctuations in the first 100 epochs, after which the model performance stabilized and improved. This study's effectiveness was validated by comparing the proposed RT-DETR method against other advanced object detection models, including Faster R-CNN, SSD, YOLOv6, YOLOv7, YOLOv8, and RDE-DETR. These models were trained and tested, and their performance metrics are detailed in Table 6. The RDE-DETR model demonstrated superior performance in detecting wheat ears, with precision, recall, F1 score, and mAP reaching 97.3%, 97.0%, 97.4%, and 98.3%, respectively, showing respective improvements over the YOLOv8 model by 2.0, 1.4, 2.0, and 3.4 percentage points. It also showed lower miss and false detection rates and demonstrated notable speed advantages in real-time detection scenarios. RDE-DETR's efficiency in parameter count and memory usage highlighted its robustness, particularly under complex field conditions.

Compared to other models, RDE-DETR model achieved excellent false positive control (FPR = 2.04%), thanks to two architectural innovations—dynamic proportional attention: reducing background misclassification by adaptively suppressing non-ear areas in dense tree crowns; and anchor free paradigm: eliminating anchor mismatch artifacts that caused 68% of YOLOv8 false positives in occluded scenes. False negatives (FNR) mainly occur in two situations—stage transition ambiguity: 3.2% of immature ears exhibit spectral similarity with mature ears when approaching the maturity threshold; and extreme occlusion: dense tillering leads to 1.8% missed detections, although its performance is better than other models through transformer-based global context modeling.

As depicted in Figure 8, the mean average precision (mAP) performance metrics and their confidence intervals, calculated based on 95% confidence level (95% CI), for six distinct object detection models—Faster R-CNN, SSD, YOLOv8, YOLOv9, YOLOv10, and

Table 6. Network model performance comparison (%).

Models	P	R	F1	mAP	FPR	FNR	Memory usage (MB)
Faster R-CNN	92.6	88.6	90.6	90.6	5.59	11.4	100.9
SSD	78.2	90.3	83.8	90.3	16.47	9.7	88.4
YOLOv8	90.4	89.2	89.8	89.4	9.21	10.8	41.6
YOLOv9	93.2	91.6	92.4	94.0	5.14	8.4	95.3
YOLOv10	95.3	95.6	95.4	94.9	3.55	4.4	65.8
RDE-DETR	97.3	97.0	97.4	98.3	2.04	3.0	33.7

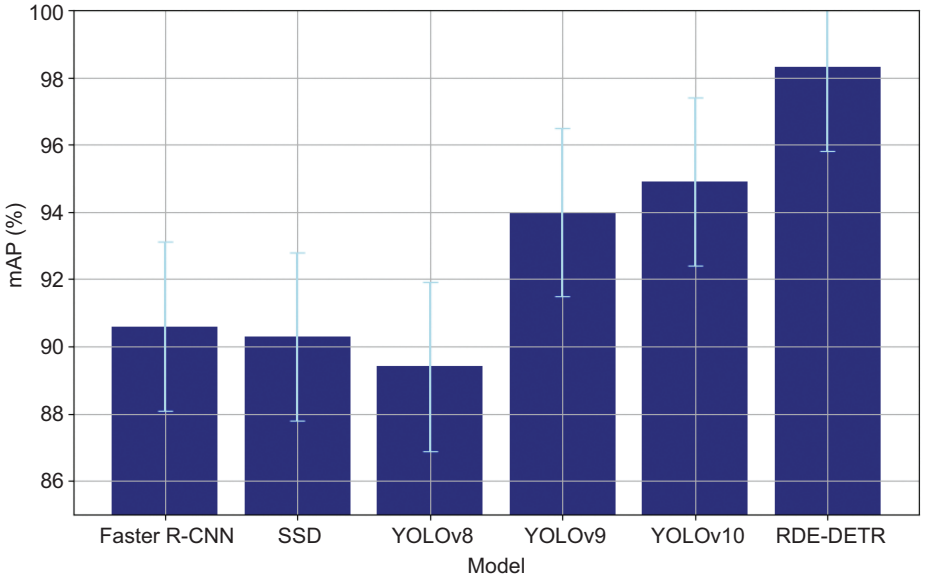

Figure 8. 95% Confidence interval (95% CI) model average accuracy value mAP.

RDE-DETR—are presented; 95% CI was derived from multiple samplings from the same distribution, indicating that approximately 95% of mAP values are within the ranges shown by error bars. The computed unified confidence interval error is approximately 1%, suggesting that mAP values for each model fluctuate about 1% above and below their estimated points.

Comparison test of different light intensities

In order to achieve all-day detection of wheat ear maturity, the dataset was collected from 08:00 am to 06:00 pm, capturing images under different lighting conditions because of varying angles of sunlight, resulting in both backlit and frontlit scenarios. The experiment utilized 120 images from these conditions to test network models under six different lighting scenarios. As shown in Table 7 and Figure 9, under frontlit conditions, the RDE-DETR model achieved a precision (P) of 97.3%, a recall (R) rate of 97.7%, an F1 score of 97.6%, and an mAP of 97.3%, outperforming the other five models in all metrics. Under shadowed conditions, RDE-DETR's precision was 95.6%, which was 7.7, 6.6, 7.1, 7.5, 3.8, and 0.1 percentage points higher than that of Faster R-CNN, SSD, YOLOv8, YOLOv9, and YOLOv10, respectively. Variation in lighting increased the difficulty of extracting wheat ear features, notably affecting the accuracy of the SSD model. This effectively addressed the challenge of detecting wheat ears against dark backgrounds, demonstrating strong robustness of the RDE-DETR detection model. In Table 7, F, B, and T represent Front Light, Back Light, and Testing time, respectively.

Comparison results of different smoothness levels

In order to verify the generalization performance of the proposed RDE-DETR model, an experiment was conducted using wheat ear images with various levels of smoothness and multiple targets. A total of 101 images in noisy and smooth conditions were selected for testing. The models used for training and detection included Faster R-CNN, SSD, YOLOv8, YOLOv9, YOLOv10, and RDE-DETR, with a counting prediction performed on four wheat ear images for each of the six detection models. The comparative results of model performance on images with different smoothness levels are summarized in Table 8. S, N, and T in Table 8 represent Smooth, Noise addition, and Testing time, respectively.

According to the comparative data in Figure 8 and Table 8, the RDE-DETR model demonstrated higher precision (P), recall (R), F1 score, and mAP@0.5 (IoU threshold of 0.5) values under both noisy and smoothness conditions, compared to other five models. Specifically, under noisy conditions, Faster R-CNN and SSD models achieved P, R, F1, and mAP@0.5 values of 88.7, 89.3, 86.0, 89.7, and 79.5, 88.6, 83.8, 90.5, respectively. Under noisy conditions, the performance of other models is relatively low because of the unclear wheat ear features in the fuzzy state, resulting in poor detection performance.

Conclusion

(1) In the proposed RDE-DETR model integrates ResNet18 as the backbone, DSConv layers, and an enhanced multi-head self-attention mechanism, achieving superior wheat spike detection accuracy. Trained on 3,370 images, it attains a precision of 97.8%, an F1 score of 97.4%, and an mAP@0.5 (IoU threshold of 0.5) of 98.3%, outperforming Faster R-CNN and YOLOv8-v10 by 2.5–19.6% across metrics. The model demonstrates robustness under challenging conditions (backlight/strong light/noisy), maintaining mAP@0.5 ≥ 97.1% while operating with a lightweight footprint (33.7 MB). Its compatibility with multi-source data (smartphone/camera) underscores potential for UAV-based real-time agricultural monitoring. In the RDE-DETR model, the backbone network was replaced with ResNet18, and convolutions were substituted with DSConv alongside an improved multi-head self-attention mechanism. The model was trained and tested using a dataset of 3,370 images of

Table 7. Different lighting detection results (%).

	Models											
	Faster R-CNN		SSD		YOLOv8		YOLOv9		YOLOv10		RDE-DETR	
	F	B	F	B	F	B	F	B	Ft	B	F	B
P	94.7	87.7	82.4	80.5	91.2	88.0	93.4	91.9	93.9	93.3	96.8	94.7
R	91.4	89.4	81.1	88.6	91.8	89.6	92.3	92.4	95.8	95.6	97.7	97.5
F1	92.6	85.8	86.5	83.6	91.4	88.8	92.7	92.3	94.3	94.8	97.6	96.3
mAP	92.1	89.6	91.2	90.7	92.9	90.2	94.9	93.5	96.8	97.2	98.9	97.1
T (ms)	135.9		126.2		55.3		43.2		33.1		29.5	

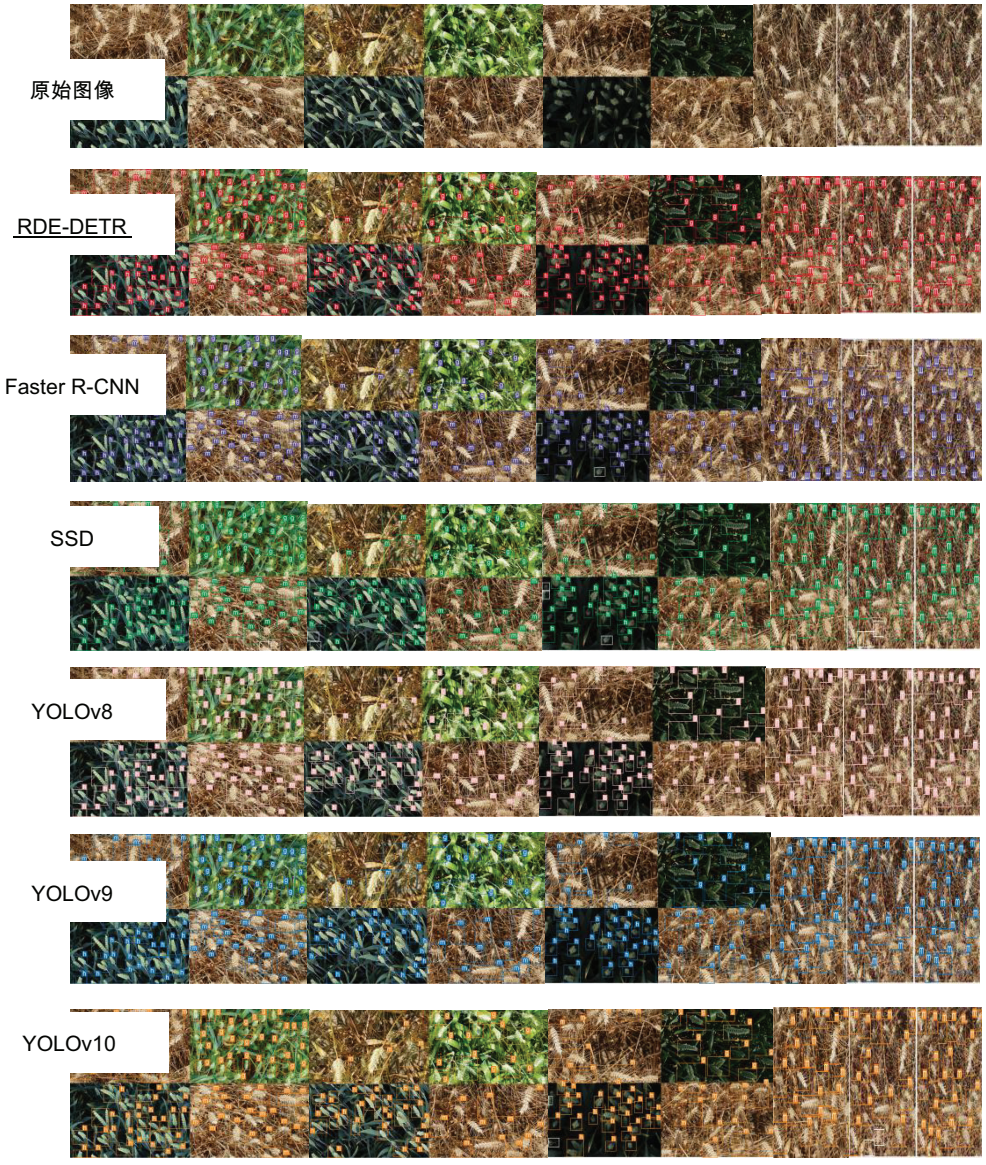

Figure 9. Detection results of different target models.

Table 8. Different smoothness level results (%).

	Models											
	Faster R-CNN		SSD		YOLOv8		YOLOv9		YOLOv10		RDE-DETR	
	S	N	S	N	S	N	S	N	S	N	S	N
P	94.6	88.7	82.6	79.5	90.3	87.9	93.1	92.6	94.6	94.3	97.4	95.8
R	90.4	89.3	91.0	88.6	91.2	89.6	92.1	92.4	96.1	95.6	97.6	97.0
F1	92.5	86.0	86.6	83.8	91.6	88.7	92.6	92.5	95.8	94.9	97.5	96.3
mAP	92.1	89.7	91.1	90.5	92.3	90.1	94.7	93.6	96.9	97.1	98.1	97.8
T (ms)	136.1		125.8		54.9		45.3		32.6		28.9	

wheat spikes, achieving an optimized network model with a precision of 97.8%, a recall of 97.0%, an F1 score of 97.4%, and an mAP@0.5 of 98.3%. Compared with other detection models, such as Faster R-CNN, SSD, YOLOv8, YOLOv9, and YOLOv10, the RDE-DETR model demonstrated superior performance, showing improvements in precision by 5.6, 19.6, 7.4, 4.6, and 2.5 percentage points; F1 scores by 7.7, 8.0, 8.9, 5.0, and 2.0 percentage points; and mAP@0.5 by 7.7, 8.0, 9.9, 4.3, and 3.4 percentage points, respectively.

(2) Despite the progress made, RDE-ETR still has limitations in high-density crop environments and extreme weather conditions. Future research must prioritize integrating state-of-the-art attention mechanisms (such as dynamic sparse attention) and lightweight backbone networks (such as MobileNetV4), utilizing multimodal fusion of near-infrared spectra to enhance generalization ability in complex on-site scenes. In addition, the framework can be extended to disease and pest identification and multi-crop monitoring, positioning it as a scalable solution for large-scale precision agriculture applications.

Data Availability Statement

The datasets presented in this article are not readily available because the data are part of an ongoing study.

Author Contributions

Mingyue Yan: situational analysis, data management, original draft writing, review, and editing. Jingfa Yao and Guifa Teng: supervision. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declared no conflict of interest.

Funding

This research was funded by the National Natural Science Foundation of China (Grant No. U20A20180), the Agricultural Science and Technology Achievement Transformation Fund Project of Hebei Province (Grant No. V1705309944504), Baoding Philosophy and Social Sciences Planning Project (Grant No. 2024072), Hebei Province Key Research Program Project (Grant No. 21327405D), and China University Industry University Research Innovation Fund (Grant No. 2021LDA10005).

References

Chen, J., Ji, C., Zhang, J., *et al.* 2024. A method for multi-target segmentation of bud-stage apple trees based on improved YOLOv8. Computers and Electronics in Agriculture 220: 108876. https://doi.org/10.1016/j.compag.2024.108876

Du, Y. 2019. Research on Monitoring the Filling Process and Maturity of Winter Wheat Based on Multi source Remote Sensing Data. Yangzhou University, Yangzhou, China. https://doi.org/10.27441/d.cnki.gyzdu.2019.000651

Helaly, R., Messaoud, S., Bouaafia, S., *et al.* 2023. DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. Signal, Image and Video Processing 17(6): 2731–2744. https://doi.org/10.1007/s11760-023-02490-6

Hu, J., Zhang, G., Shen, M., *et al.* 2024. Improved RT-DETR model for surface defect detection of pine wood. Journal of Agricultural Engineering 40(07): 210–218.

Jia, J., Fu, M., Liu, X., *et al.* 2022. Underwater object detection based on improved efficientDet. Remote Sensing 14(18): 4487. https://doi.org/10.3390/rs14184487

Jin, S., Zhang, W., Yang, P., *et al.* 2022. Spatial-spectral feature extraction of hyperspectral images for wheat seed identification. Computers and Electrical Engineering 101: 108077. https://doi.org/10.1016/j.compeleceng.2022.108077

Khaki, S., Safaei, N., Pham, H., *et al.* 2022. WheatNet: a lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. Neurocomputing 489: 78–89. https://doi.org/10.1016/j.neucom.2022.03.017

Kumar, D. and Kukreja, V. 2022. Deep learning in wheat diseases classification: a systematic review. Multimedia Tools and Applications 81(7): 10143–10187. https://doi.org/10.1007/s11042-022-12160-3

Li, S., Hu, Z., Zhaom, M., *et al.* 2021. Cascade-guided multi-scale attention network for crowd counting. Signal, Image and Video Processing 15: 1663–1670. https://doi.org/10.1007/s11760-021-01903-8

Li, Y., Yang, B., Zhou, S., *et al.* 2022. Identification lodging degree of wheat using point cloud data and convolutional neural network. Frontiers in Plant Science 13: 968479. https://doi.org/10.3389/fpls.2022.968479

Liu, Q., Liu, Y. and Lin, D. 2023. Revolutionizing target detection in intelligent traffic systems: Yolov8-snakevision. Electronics 12(24): 4970. https://doi.org/10.3390/electronics12244970

Liu, M., Wang, H., Du, L., *et al.* 2024. Bearing-DETR: a lightweight deep learning model for bearing defect detection based on RT-DETR[J]. Sensors 24(13): 4262. https://doi.org/10.3390/s24134262

Ma, J., Li, Y., Du, K., *et al.* 2020. Segmenting ears of winter wheat at flowering stage using digital images and deep learning. Computers and Electronics in Agriculture 168: 105159. https://doi.org/10.1016/j.compag.2019.105159

Meng, J., Wu, B., Du, X., *et al.* 2011. Remote sensing prediction of winter wheat maturity based on HJ-1A/1B data. Journal of Agricultural Engineering 27(3): 225–230

Su, Y. 2011. Research on Grain Moisture Detection and Impurity and Imperfect Grain Identification Method Based on Machine Vision and Hyperspectral Image Technology. Zhejiang University, Hangzhou, China.

Wang, X., Huang, J., Feng, Q., *et al.* 2020. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sensing 12(11): 1744. https://doi.org/10.3390/rs12111744

Xu, L., Cao, B., Zhao, F., *et al.* 2023. Wheat leaf disease identification based on deep learning algorithms. Physiological and Molecular Plant Pathology 123: 101940. https://doi.org/10.1016/j.pmpp.2022.101940

Yang, B., Zhu, Y., Zhou S. 2021. Accurate wheat lodging extraction from multi-channel UAV images using a lightweight network model. Sensors 21(20). https://doi.org/10.3390/s21206826

Yao, J., Song, B., Chen, X., *et al.* 2024. Pine-YOLO: a method for detecting pine wilt disease in unmanned aerial vehicle remote sensing images. Forests 15(5): 737. https://doi.org/10.3390/f15050737

Yin, M., He, S., Soomro, T.A., *et al.* 2023. Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. Expert Systems with Applications 226: 120080. https://doi.org/10.1016/j.eswa.2023.120080

Zang, Y., Yu, Z., Xu, K., *et al.* 2022. Multi-span long-haul fiber transmission model based on cascaded neural networks with multi-head attention mechanism. Journal of Lightwave Technology 40(19): 6347–6358. https://doi.org/10.1109/JLT.2022.3195949

Zhang, T., Yang, Z., Xu, Z., *et al.* 2022. Wheat yellow rust severity detection by efficient DF-UNet and UAV multispectral imagery. IEEE Sensors Journal 22(9): 9057–9068. https://doi.org/10.1109/JSEN.2022.3156097

Zhong, L., Hu, L., Zhou, H., *et al.* 2019. Deep learning based winter wheat mapping using statistical data as ground references in Kansas and northern Texas, US. Remote Sensing of Environment 233: 111411. https://doi.org/10.1016/j.rse.2019.111411

Zhong, C., Xiong, F., Pan, S., *et al.* 2023. Hierarchical attention neural network for information cascade prediction. Information Sciences 622: 1109–1127. https://doi.org/10.1016/j.ins.2022.11.163

Zhu, X., He, Z., Zhao, L., *et al.* 2022. A cascade attention based facial expression recognition network by fusing multi-scale spatio-temporal features. Sensors 22(4): 1350. https://doi.org/10.3390/s22041350

Zhu, M. and Kong, E. 2024. Multi-scale fusion uncrewed aerial vehicle detection based on RT-DETR. Electronics 13(8): 1489. https://doi.org/10.3390/electronics13081489