Automatically Detect and Count Wheat Spikes in Complex Fields by the Improved RT-DETR

Junxiong Wang, Yaqin Zhao, Hao Ma, Ying Pu

Abstract—**Accurate wheat yield prediction critically depends on the precise detection and counting of wheat spikes. Although machine vision has emerged as a promising solution for this task, challenges such as dense spike distributions, morphological variability, and interference from weeds continue to hinder the performance of existing methods. To address these issues, this study presents an enhanced version of the multi-object detection model RT-DETR, tailored for automated wheat spike detection and counting in images. A novel feature extraction module, FasterSEV2, is proposed by integrating the SENetV2 attention mechanism into the FasterNet architecture, replacing the original backbone convolution of RT-DETR. This modification significantly reduces redundancy and computational complexity. Furthermore, the RT-DETR encoder is replaced with RepGFPN, which effectively captures both local and global features of wheat spikes, thereby improving robustness to variations in size and morphology. The improved RT-DETR is also integrated with DeepSORT, enabling accurate tracking and counting of wheat spikes. To facilitate evaluation under challenging conditions, a meticulously annotated dataset of wheat spikes is constructed. Experimental results indicate that the proposed method achieves a mean Average Precision of 94.5%, representing a 1.1% improvement over the original model. These findings demonstrate the effectiveness and practical applicability of the proposed approach for automated wheat spike counting in the context of precision agriculture.**

Index Terms—**Wheat spike detection, precision agriculture, automatic counting, improved RT-DETR.**

I. Introduction

As one of the most important food crops worldwide, wheat plays a critical role in ensuring food security and promoting the sustainable development of agriculture [1]. Accurate counting of wheat spikes is vital for yield prediction, and can provide essential data for the adjustment of planting density in subsequent growing seasons. However, manual spike counting is impractical due to the small size and dense spatial distribution of wheat in field environments. With advancements in computer vision, the deployment of fixed cameras and unmanned aerial vehicles (UAVs) for monitoring crop growth and maturity has become a central focus in precision agriculture research. Recent studies have explored image-based wheat spike counting methods and their application in agricultural settings [2][3].

Traditional approaches to automatic wheat spike counting typically rely on the extraction of handcrafted features such as shape [4], texture [5], and color [6], followed by the use of trained classifiers to identify and count spikes [7][8][9]. Although these methods may achieve acceptable performance under controlled conditions, their effectiveness diminishes in diverse field environments and across varying growth stages, where visual differences substantially increase the rates of both missed and false detections. Moreover, the selection of appropriate features often requires extensive empirical experimentation and analysis, limiting the scalability and robustness of such methods [10]. In recent years, advances in deep learning have significantly improved computer vision performance, with convolutional neural network (CNN) [11]-based object detection frameworks—such as the YOLO [12][13][14] and R-CNN [15][16][17] families—receiving substantial attention for wheat spike recognition. Wu et al. [18] achieved automated spike counting by optimizing the YOLOv7 model, while Li et al. [19] enhanced detection efficiency by integrating Faster R-CNN with RetinaNet. Despite the real-time inference capabilities of YOLO-based models, these architectures continue to face notable challenges in accurately detecting densely distributed wheat spikes characterized by chromatic and morphological variability. In particular, YOLO-based approaches often exhibit reduced performance under partial occlusion and show limited robustness in distinguishing spikes from visually similar background elements, such as weeds. Although R-CNN variants typically offer higher detection precision, they impose considerable computational overhead due to their region proposal mechanisms, especially when applied to high-resolution agricultural imagery with complex backgrounds. Alternative approaches have also been explored. Sadeghi-Tehran et al. [20] applied superpixel segmentation using linear iterative clustering (LIC) in conjunction with CNN-based feature modeling, while Wang et al. [21] proposed a fully convolutional network (FCN) architecture that integrates Harris corner-based feature extraction. Despite these advancements, existing methods consistently encounter high error rates under challenging conditions such as spike overlap, weed interference, and variations in illumination—factors that pose significant obstacles to the reliable deployment of these models in real-world agricultural settings.

Manuscript received March 31, 2025; revised Jun 21, 2025.

Junxiong Wang is a postgraduate student of the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, Jiangsu, China (e-mail: 14779736243@163.com).

Yaqin Zhao is a professor at the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, Jiangsu, China. (Corresponding author, e-mail: yaqinzhao@163.com).

Hao Ma is a postgraduate student of the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, Jiangsu, China (e-mail: 13961327323@163.com).

Ying Pu is a postgraduate student of the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, Jiangsu, China (e-mail: py17372725766@163.com).

Fig 1 The examples of wheat spike images

TABLE I
DIVISION OF THE DATASET

	Training set	Validation set	Test set	Total
GWHD	3093	368	840	4301
Self-made	406	102	206	714
Total	3499	408	1046	5015

Accurately counting wheat spikes across an entire field in agricultural monitoring is crucial, necessitating methods that prevent duplication or omission. Video-based approaches are preferred over static image-based techniques for their comprehensive and dynamic tracking capabilities over large areas. These video-based tracking methods, previously utilized for fruit counting in diverse crops, have been increasingly employed in tracking wheat spike. Recent studies have predominantly utilized YOLO-based models as object detectors in conjunction with Deep SORT as the tracking algorithm for spike tracking and counting in wheat.

In order to enhance applicability in agricultural settings, this study combines Deep SORT with an upgraded version of the Real-time Object Detection model (RT-DETR) known as DETRs Beat YOLOs. RT-DETR, renowned for its efficiency in multi-object detection tasks, strikes a favorable balance between speed and accuracy. Nevertheless, deploying the original RT-DETR in real-world agricultural conditions is hindered by challenges like dense spike distribution, weed interference, and phenotypic variations across growth stages. To address these challenges, two targeted enhancements to the RT-DETR architecture are proposed.

(1) Firstly, a novel feature extraction module named FasterSEV2 is introduced to capture global information and local details of wheat spikes efficiently while minimizing computational burden. FasterSEV2 integrates the SENetV2 into the FasterNet architecture. (2) Secondly, the backbone block of RT-DETR is replaced with the FasterSEV2 module to decrease false and missed detection rates, especially in scenarios involving occluded wheat spikes and weed interference. (3) Additionally, the RT-DETR encoder is substituted with the RepGFPN module, a multi-scale fusion module, to integrate detailed low-level features with high-level semantic information. This modification broadens the model's adaptability to diverse wheat fields characterized by significant variations in spike colors and shapes.

II. EXPERIMENTAL DATA

The publicly accessible Global Wheat Head Detection Dataset (GWHD) [29] is employed to ensure impartial training and assessment of the proposed model. This dataset comprises information from various wheat cultivation areas worldwide, encompassing over 3,000 high-resolution images, each with a resolution of 1024 × 1024 pixels. Standardization has been applied to all original images, involving center-preserving cropping to 1024 × 1024 pixel regions.

To enhance the dataset's generalization capacity and adapt it to practical applications in wheat fields, we construct a dataset comprising challenging images of wheat spikes. Image acquisition was conducted using a fixed Sony α7R IV full-frame camera, equipped with a 35mm prime lens, for vertical overhead shooting at a height of 1.5 meters. The high-resolution sensor (9506 × 6336 pixels) is used to capture images during three critical growth stages of wheat: the milking, wax-ripening, and full-ripening stages. The data includes some challenging scenes, such as dense wheat spikes, changes in illumination, and weed interference.

A significant amount of manual labor was dedicated to annotating wheat spike regions in all images to preprocess the data for training deep learning models for wheat spike detection. Fig 1 shows typical samples from the two datasets: the first row depicts samples from the GWHD dataset, and the second row shows samples from the self-made dataset. These samples illustrate the variety and complexity of the wheat spike scenes. The dataset's comprehensive composition and statistics are outlined in Table 1.

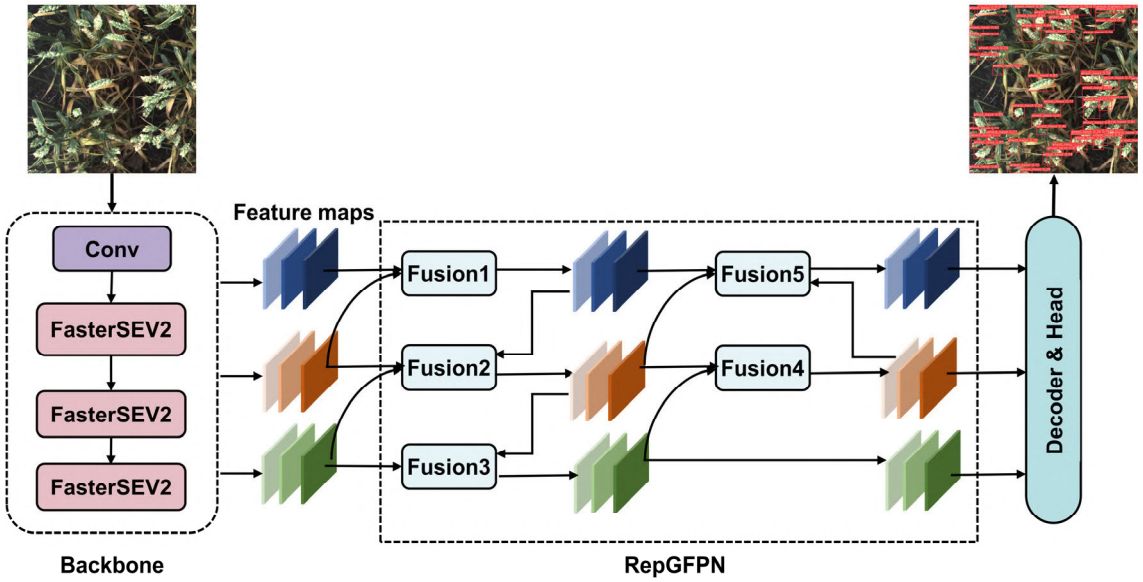


Fig 2 Overall framework of the wheat spike counting model

III. Mode

This study introduces a novel Transformer-based network designed for wheat spike counting by enhancing RT-DETR, a high-performing multi-object detection model. The network architecture, depicted in Fig 2 consists of three main components: the Backbone, RepGFPN, and Decoder. Initially, the SENetV2 attention mechanism is incorporated into the partial convolutional structure of FasterNet to create a new backbone module named FasterSEV2. The wheat images are passed through the improved backbone network for feature extraction. The feature maps from the last three layers are then fed into RepGFPN for feature fusion. Finally, the decoder completes the detection process. This module replaces the original backbone in RT-DETR, effectively reducing computational redundancy while improving the network's capacity to capture both local details and global contextual information of wheat spikes. This adjustment helps alleviate false positives and missed detections, particularly in scenarios involving partial occlusion or interference from weeds. Subsequently, the encoder of RT-DETR is substituted with the RepGFPN module to facilitate efficient multi-scale feature fusion, thereby enhancing detection performance across various wheat maturity stages, such as the milk, dough, and full ripening phases. Finally, the integrated feature maps are inputted into the decoder for prediction.

A. Wheat spike feature extraction module FasterSEV2

In wheat fields, spikes are typically small and densely distributed, often leading to occlusion by surrounding foliage. Moreover, the visual similarity between weeds and wheat spikes presents additional challenges, frequently resulting in false positives and missed detections. In real-world agricultural scenarios, ensuring high detection efficiency is essential for the practical deployment of the model.

The Backbone of RT-DETR utilizes the conventional convolutional architecture ResNet18. In ResNet18, convolution is applied across all channels of the input feature map, resulting in high computational complexity. The ResNet18 is replaced with FasterNet, which is better suited to meet the real-time requirements of wheat spike counting. Partial convolution (PConv), a core component of FasterNet, applies a regular convolution to only a part of the input channels for spatial feature extraction and leaves the remaining channels untouched, thereby significantly reducing unnecessary computation and memory access. Although FasterNet excels in inference speed and computational efficiency, it encounters challenges in detecting wheat spikes, particularly in the face of weed interference. To address these issues, the attention module SENetV2 is incorporated into FasterNet, resulting in the creation of a new feature extraction module termed FasterSEV2. This integration enhances the model's ability to capture wheat spike features, thereby improving its performance in wheat spike detection tasks.

The architecture of FasterSEV2 is shown in Fig 3 Firstly, partial convolution (PConv) operates on the channels of the input feature map, where c represents the total number of channels. In contrast to traditional convolution that processes all c channels, PConv significantly reduces redundant computations and memory access. Following PConv, two pointwise convolution (PWConv) operations are applied. The PConv+PWConv structure effectively integrates information across all channels, creating a T-shaped convolution architecture that emphasizes central features [30]. Then, the SENetV2 module is incorporated into the last convolutional layer of FasterNet. SENetV2 enhances feature expression by introducing an aggregated dense layer that employs a multi-branch convolutional architecture. Specifically, after standard convolution, the output undergoes an activation operation, which is realized through two fully connected layers that utilize two 1×1 convolutional layers to compute channel-wise weights. The aggregated result is then passed through another convolution, producing a new feature map. This process strengthens the model's ability to capture complex inter-channel dependencies, improving the extraction of both local details and global features of the wheat spikes.

This integration significantly reduces false positives and missed detection, particularly in cases of partial occlusion of wheat spikes or weed interference.

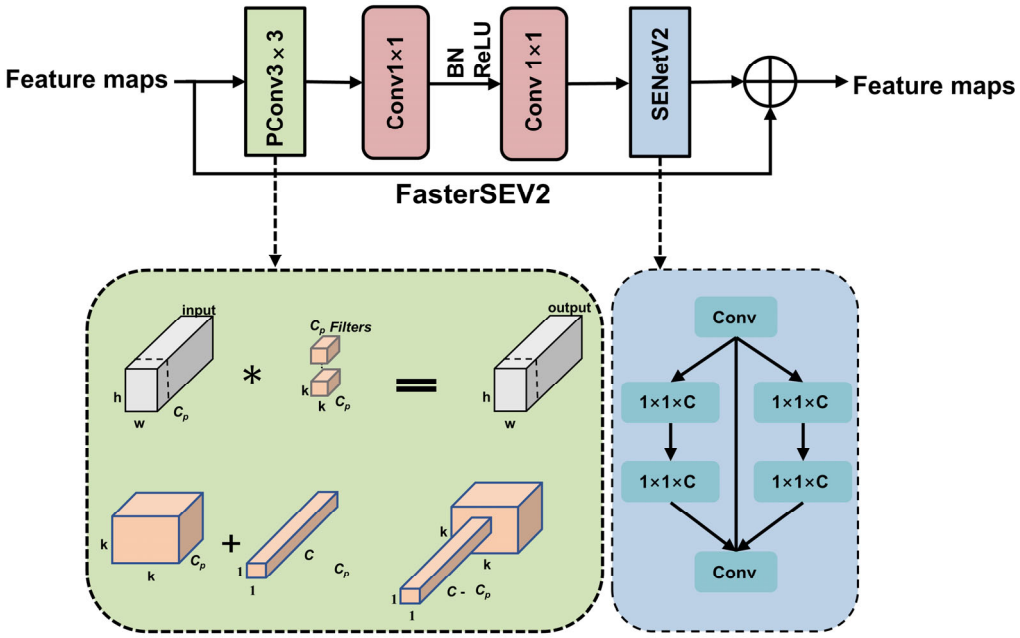


Fig 3 Overall architecture of FasterSEV2

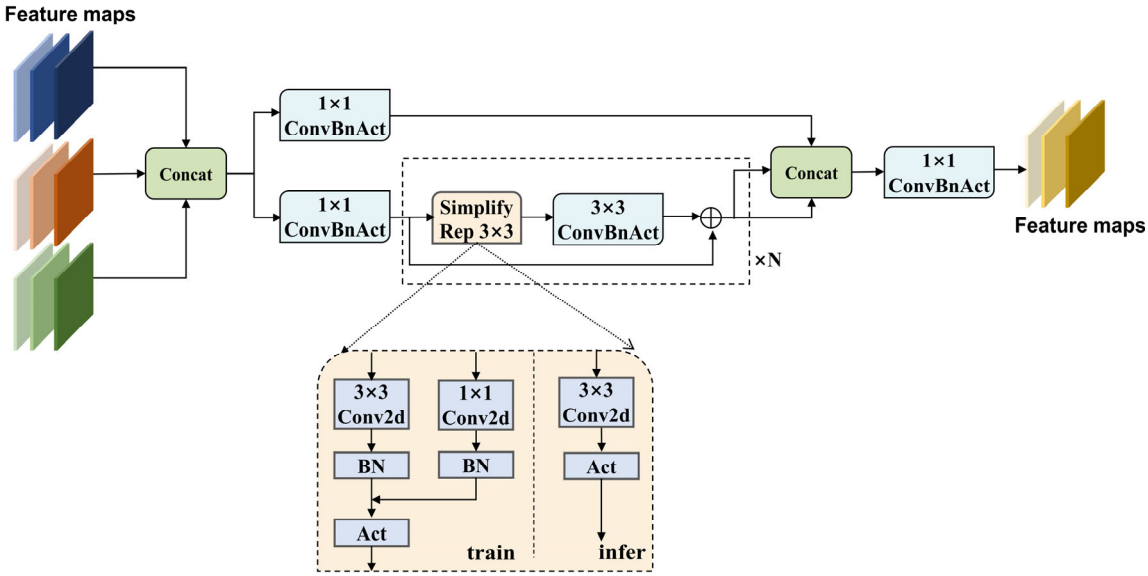


Fig 4 Structure of the fusion module

B. Multi-scale feature fusion of wheat spikes based on RepGFPN

Significant variations in the color, size, and morphology of wheat spikes occur across different maturity stages (i.e., milk, wax, and full maturity). Concurrently, the color of the straw and leaves also evolves with crop development. For instance, during the milk stage, the spikes, straw, and leaves exhibit a gradual transition from green to yellow. By full maturity, these components adopt a golden hue, and the grain size of the wheat spikes becomes noticeably smaller than in earlier stages. These dynamic visual changes present substantial challenges for accurate wheat spike detection.

The encoder in the original RT-DETR model predominantly focuses on large and prominent targets during feature extraction, limiting its ability to capture the fine-grained details of smaller wheat spikes. To address this limitation and enhance the model's capacity to represent the distinctive characteristics of wheat spikes, the original encoder is replaced with the Re-parameterized Generalized Feature Pyramid Network (RepGFPN). RepGFPN employs a feature pyramid structure to extract multi-scale features and integrates a heavily parameterized convolutional design along with a multi-scale feature aggregation mechanism to effectively fuse information across different feature levels. This enhancement significantly improves the model's ability to detect wheat spikes of varying sizes and shapes under complex field conditions.

The re-parameterized convolution simplifies complex multi-branch convolutional structures into a more efficient single-branch convolution during inference. In the training phase, multiple convolutional branches are employed to capture fine-grained features of wheat spikes across different scales and orientations, enabling the network to learn more diverse and informative representations. Branches with larger convolutional kernels are responsible for capturing global shape and positional cues, whereas branches with smaller kernels focus on edge details and

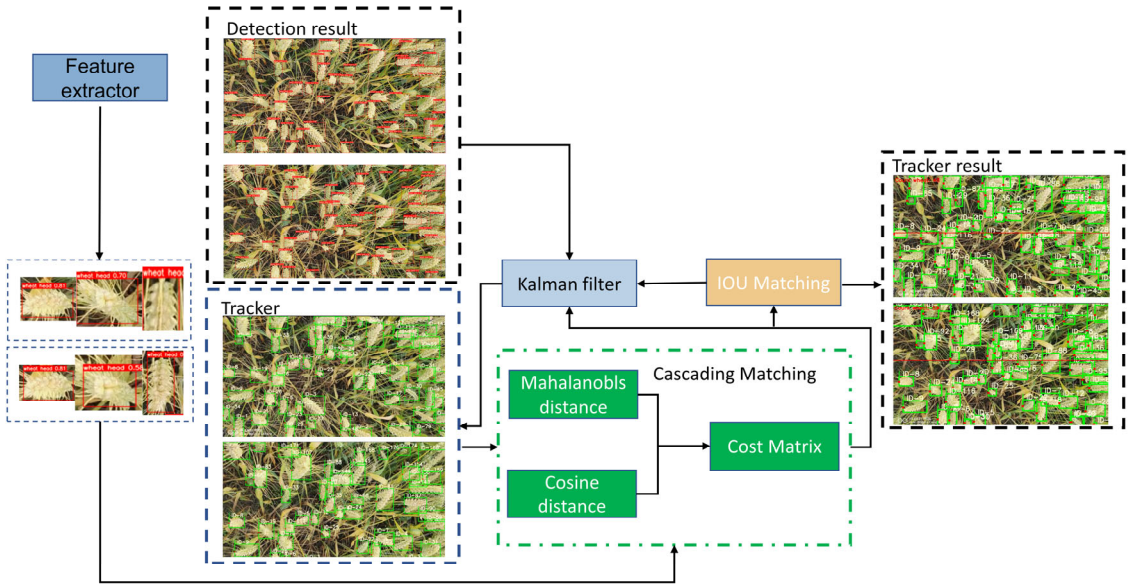


Fig 5 The framework of object tracking

textural patterns. During inference, these branches are merged into a single, re-parameterized convolutional operation, which significantly reduces computational complexity while preserving the model's feature extraction capabilities. Additionally, the multi-scale feature aggregation mechanism is designed to integrate features from multiple spatial resolutions effectively. RepGFPN fuses low-level spatial details with high-level semantic representations to construct a more comprehensive and robust feature map. By attending to features across multiple scales, RepGFPN enhances the model's ability to accurately localize and delineate wheat spikes of varying sizes and shapes.

The Fusion block, as illustrated in Fig 4 integrates multiple 1×1 convolutional layers with re-parameterized convolutions to enhance multi-scale feature representation. Initially, feature maps are processed in parallel branches. In the upper branch, a 1×1 convolution is directly applied, and its output is forwarded to the concatenation stage. In the lower branch, a sequence of 1×1 followed by 3×3 convolutions is employed to extract more complex spatial features. Both branches aim to reduce the dimensionality of the input feature maps. Subsequently, the outputs of the two branches are concatenated and passed through an additional 1×1 convolution to fuse the information, resulting in enhanced multi-scale feature representations.

C. Wheat spike Tracking and Counting

During the capture of wheat spikes in field conditions, instances of missed frames or overlapping regions between individual images may occur. As a result, relying solely on object detection can lead to duplicate or missed counts. To address this issue, the integration of object tracking algorithms is proposed. Multi-Object Tracking (MOT) aims to track multiple targets across consecutive video frames by assigning each detected object a unique identity (ID), thereby ensuring consistent identification and tracking throughout the video sequence. Among the existing tracking algorithms, DeepSORT has been widely adopted for spike counting tasks. Previous studies have demonstrated that DeepSORT effectively reduces duplicate and missed counts by combining spatial and appearance features. Accordingly, DeepSORT is employed in this study as the tracking algorithm for wheat spikes. The proposed wheat spike tracking pipeline is illustrated in Fig 5. Specifically, the improved RT-DETR model is first applied to perform object detection and generate bounding boxes for wheat spikes in each frame. Subsequently, DeepSORT is used to predict object trajectories and associate detections across frames, assigning unique IDs to each tracked spike. This approach facilitates real-time, accurate counting and robust tracking across sequential frames.

Due to the partial occlusion of wheat spikes and camera motion, individual spikes may repeatedly appear within the detection range across different frames, leading to ID loss or switching. This results in duplicate or missed counts. To address these issues, a virtual fixed counting line is established in each video frame, and a multi-frame cross-line counting strategy combined with a multi-frame confirmation mechanism is employed to ensure accurate spike counting [24]. As illustrated in Fig 6 ,a virtual counting line is placed in the central region of the video frame to delineate two counting zones, namely Area A and Area B. A spike is considered to have crossed the virtual line and entered Area A if its center pixel coordinates in Area B are greater than or equal to the pixel coordinates of the line in the current frame, and subsequently, the center coordinates fall below the line in future frames. When a spike is partially located in both areas (i.e., straddling the virtual line), the count is not immediately incremented. Instead, the spike is continuously tracked across successive frames, and the count is increased by one only when the entire spike has completely crossed the line into Area A. This approach reduces false counts caused by partial visibility or motion blur and enhances counting precision.

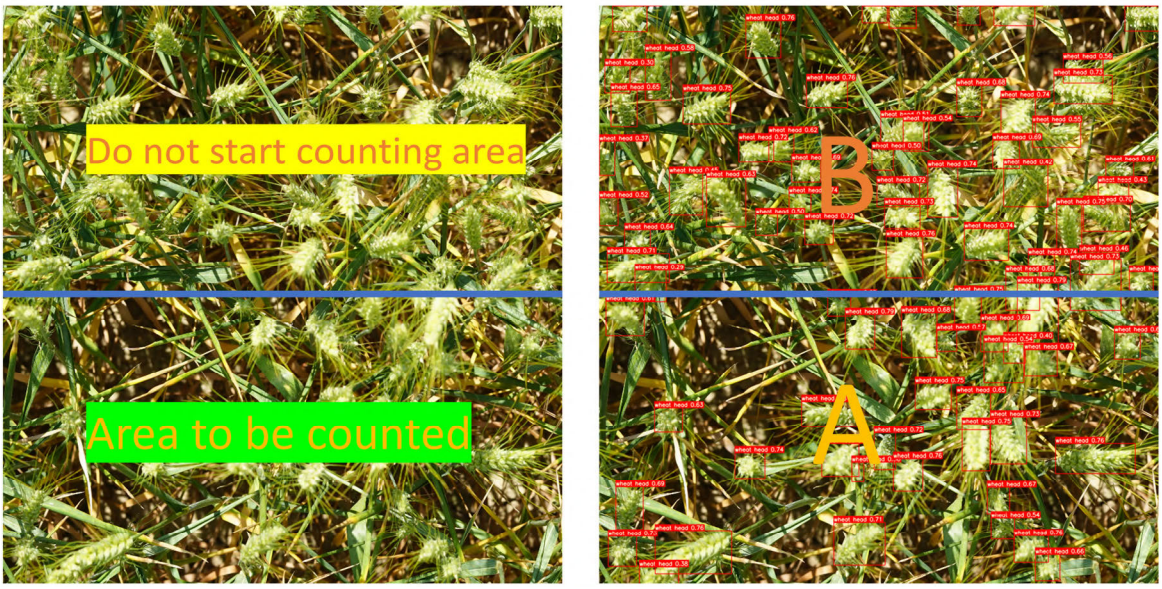


Fig 6 Multi-frame Cross-line Counting Diagram

IV Results & Discussion

A. Experiment Details

All experiments were conducted on a Windows 11 operating system, utilizing an Intel® Core i7-13700H processor and an NVIDIA GeForce RTX 4060 Laptop GPU. The software environment comprised Python 3.8 as the programming language, PyTorch 11.1 as the deep learning framework, and CUDA 11.2 for GPU-accelerated computation. The training parameters were set as follows: an initial learning rate (lr) of 0.0001, a momentum of 0.9, and a weight decay of 0.0001.

To comprehensively assess the performance of the proposed model, several widely used evaluation metrics were employed, including Precision, Recall, F1-score, and mean Average Precision (mAP). Let TP denote the number of true positives (positive instances correctly identified), TN the number of true negatives (negative instances correctly identified), FP the number of false positives (negative instances incorrectly identified as positive), and FN the number of false negatives (positive instances incorrectly identified as negative). The evaluation metrics are computed as follows:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Precision measures the accuracy of the model in the positive prediction, and Recall reflects the model's ability to identify positive samples, and F1-score provides a more comprehensive assessment of the model's performance.

Mean Average Precision (mAP) refers to the average precision at different thresholds, considering both the detection accuracy and localization precision:

$$mAP = \frac{1}{n}\sum_{j=1}^{n} AP_j \quad (4)$$

where $AP = \int_0^1 p \cdot r \quad dr$ represents the area under the Precision-Recall curve for one category, and n denotes the number of categories. This study employs mAP50 to assess the model. mAP50 denotes the mAP value when the threshold of IOU (Intersection over Union) is set at 0.5, IOU is defined as:

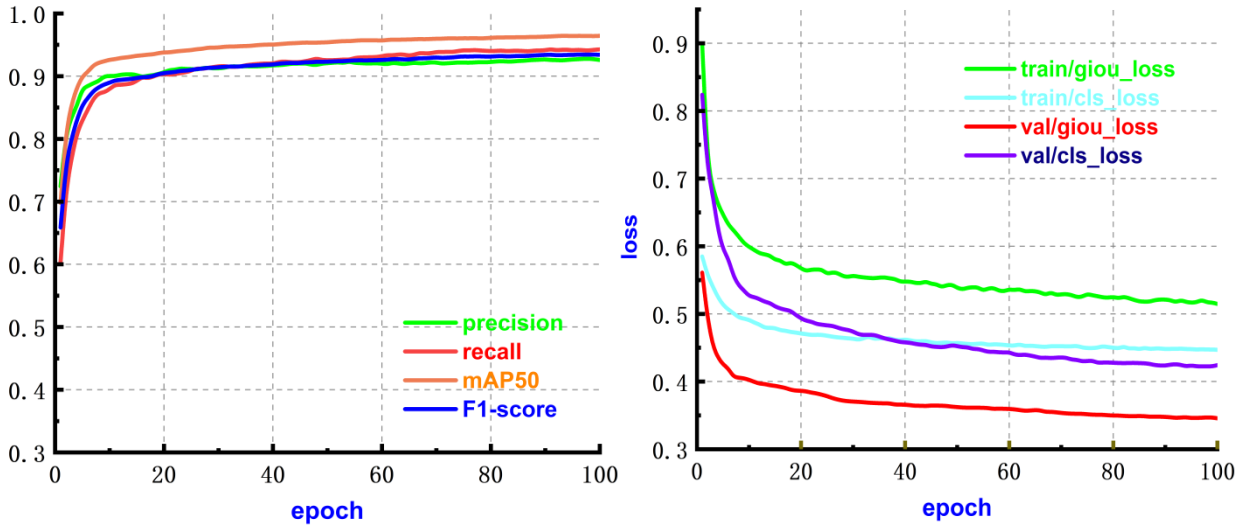
$$IOU = \frac{Area\ of\ overlap}{Area\ of\ union} \quad (5)$$

The IOU is calculated as a ratio of the area of overlap to the area of the union. The prediction results are considered correct when IOU$\geqslant$0.5.

B. Convergence of the model

The performance of the proposed model was comprehensively evaluated by analyzing the trends of key evaluation metrics and loss functions during training. Fig 7 (a) illustrates that Precisio Recall, F1-score, and mAP50 exhibited rapid increments with training epochs, followed by gradual stabilization, indicating successful model convergence. Notably, mAP50 peaked at approximately 0.97, showcasing the model's precise object localization capability. The Recall curve consistently outperformed the Precision curve, underscoring the model's effective target retrieval and wheat spike detection. The F1-score, balancing Precision and Recall, stabilized at around 0.93, affirming the model's sustained high accuracy and completeness in detection.

Fig 7 (b) depicts the variations in loss functions, specifically cls_loss and giou_loss, throughout the training and validation processes. Each of the four loss curves displayed a consistent decrease, indicative of continuous optimization throughout training. Both training and validation loss trends closely mirrored each other, suggesting strong generalization capabilities of the model and the absence of overfitting issues.

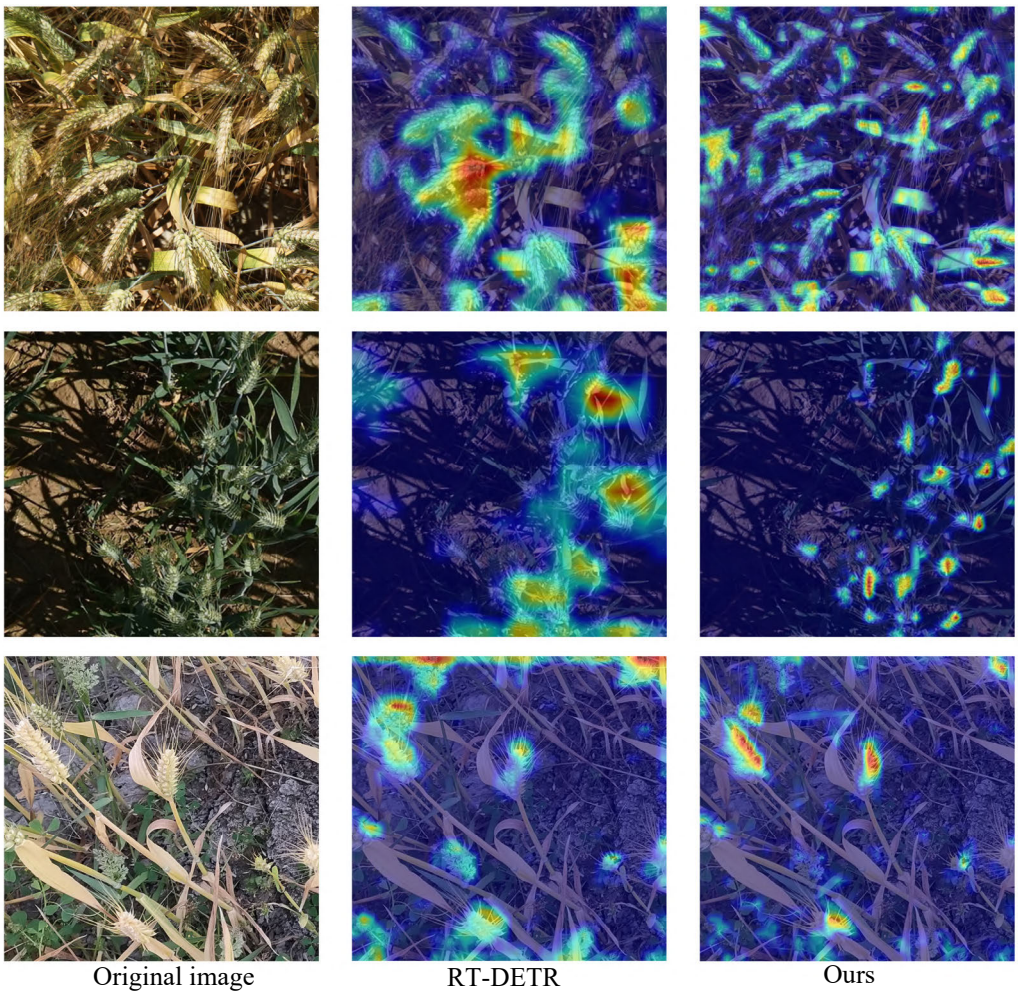


(a)evaluation metrcics (b)loss

Fig 7 Evaluation metrics and loss values with the number of iterations

TABLE II
ABLATION EXPERIMENTAL RESULTS

FasterNet	SENetV2	RepGFPN	Precision	Recall	mAP50	Param	GFlops
			90.5%	90.2%	93.4%	20.1M	58.08
√			90.9%	90.5%	93.6%	16.8M	50.66
√	√		91.2%	90.4%	94.0%	17.1M	50.67
√	√	√	91.9%	91.1%	94.5%	22.2M	64.4



Original image RT-DETR Ours

Fig 8 Heatmap of RT-DETR and the proposed method

C. *Ablation experiments*

An in-depth ablation study was conducted to analyze the effects of FasterNet, SENetV2, and RepGFPN modules on wheat spike detection performance. The baseline model, lacking these modules, achieves 90.5% Precision, 90.2% Recall, and 93.4% mAP50, with 20.1M parameters and 58.08 GFLOPs. Introducing FasterNet as a lightweight backbone reduces parameters to 16.8M and GFLOPs to 50.66, enhancing detection performance with a slight increase in Precision and Recall by .4% and .3%, respectively. This affirms FasterNet's role in accelerating inference and improving feature representation.

The integration of SENetV2 significantly boosts model accuracy, with Precision increasing to 91.2% and mAP50 to 94.%. These enhancements indicate that SENetV2 effectively recalibrates channel-wise feature responses, improving the model's capacity to differentiate between wheat spikes and background, thereby reducing false positives. While Recall exhibits minor fluctuations, the overall trend suggests enhanced classification reliability.

Upon incorporating RepGFPN, the comprehensive model achieves peak performance, with Precision, Recall, and mAP50 reaching 91.9%, 91.1%, and 94.5%, respectively. Particularly noteworthy is the .9% Recall enhancement over the SENetV2-enhanced version, underscoring RepGFPN's superior ability to capture detailed spatial information critical for detecting closely packed or partially obscured wheat spikes. The marginal increase in parameter count and computational load (22.2M, 64.4 GFLOPs) is justified by the substantial improvements in accuracy.

In conclusion, the ablation results highlight the distinct contributions of each module to the overall performance. FasterNet ensures lightweight and efficient processing, SENetV2 enhances feature selection, and RepGFPN improves spatial fusion and Recall. The amalgamation of all three components yields a well-balanced architecture that upholds high detection accuracy while maintaining computational efficiency, rendering it well-suited for practical applications in precision agriculture.

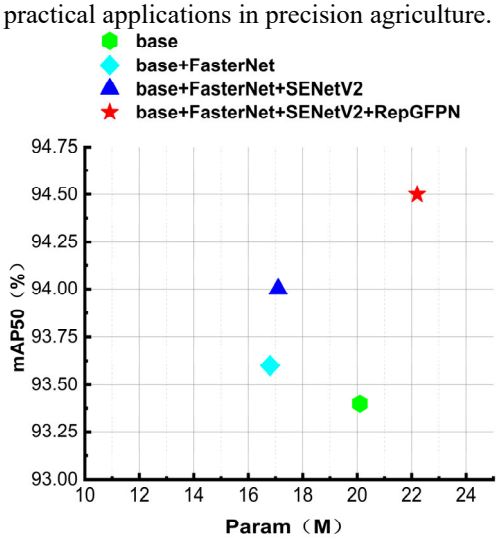


Fig 9 Parameters and mAP50 after embedding the submodule

Fig 8 displays heatmaps visualizing attention patterns from the baseline RT-DETR model and the proposed model across different challenging samples: the first row shows a sample with dense wheat spikes; the second row, a sample under low light; and the third row, a sample with highly similar (or confounding) weeds. The RT-DETR model exhibits dispersed and inaccurate attention allocation, frequently confusing background elements like weeds or soil with wheat spikes, especially in complex field settings. In contrast, the proposed model exhibits enhanced precision in identifying key wheat spike regions, with focused and consistent attention in relevant areas. The heatmaps reveal that the improved model consistently emphasizes genuine spike structures while suppressing irrelevant features like overlapping leaves or stems. This suggests that the proposed network has superior feature discrimination capabilities and resilience to visual disturbances. The refined focus not only reduces false alarms but also captures a more exhaustive range of spike instances, addressing the detection limitations seen in RT-DETR. These qualitative findings visually corroborate the quantitative enhancements observed in the ablation study and underscore the model's proficiency in interpreting dense and obscured agricultural environments.

An ablation study was conducted to evaluate the influence of various backbone networks on the model's overall performance. Only the backbone architecture was altered, while all other components remained consistent to maintain a fair comparison. The study examined five different backbones: ResNet18, FasterNet, RepViT [32], ShuffleNetV2 [33], and the newly introduced FasterSEV2. The comparative outcomes are presented in Table 3. While RepViT and ShuffleNetV2 excel in lightweight design, their performance is hindered by inferior feature representation capabilities. FasterNet, on the other hand, outperforms ResNet18 in accuracy while utilizing fewer parameters, thanks to the integration of partial convolution (PConv) for improved local feature extraction.

The integration of the SENetV2 attention mechanism into the FasterNet architecture, known as FasterSEV2, enhances the model's detection capabilities, yielding a notable improvement in performance. Notably, FasterSEV2 achieves an outstanding mAP50 score of 94.5%, demonstrating superior precision (91.9%) and recall (91.1%). These findings underscore the ability of FasterSEV2 to effectively optimize the trade-off between detection accuracy and computational efficiency, rendering it well-suited for detecting wheat spikes in intricate agricultural settings. Despite a modest increase in model parameters and computational workload, the enhanced accuracy validates the efficacy of the proposed backbone design.

D. *Performance evaluation*

A comparative evaluation was conducted to validate the effectiveness of the proposed model against state-of-the-art object detection frameworks, namely Faster R-CNN, SSD, YOLOv7, and RT-DETR. The performance results are summarized in Table 4, demonstrating the superior performance of the proposed model across all evaluation metrics. Notably, the proposed model achieves a leading mAP50 score of 94.5%, surpassing RT-DETR by 1.1 percentage points and YOLOv7 by 3.7 percentage points. Regarding Precision and Recall, the proposed model demonstrates values of 91.9% and 91.1%, respectively, indicating its robustness in accurately identifying true

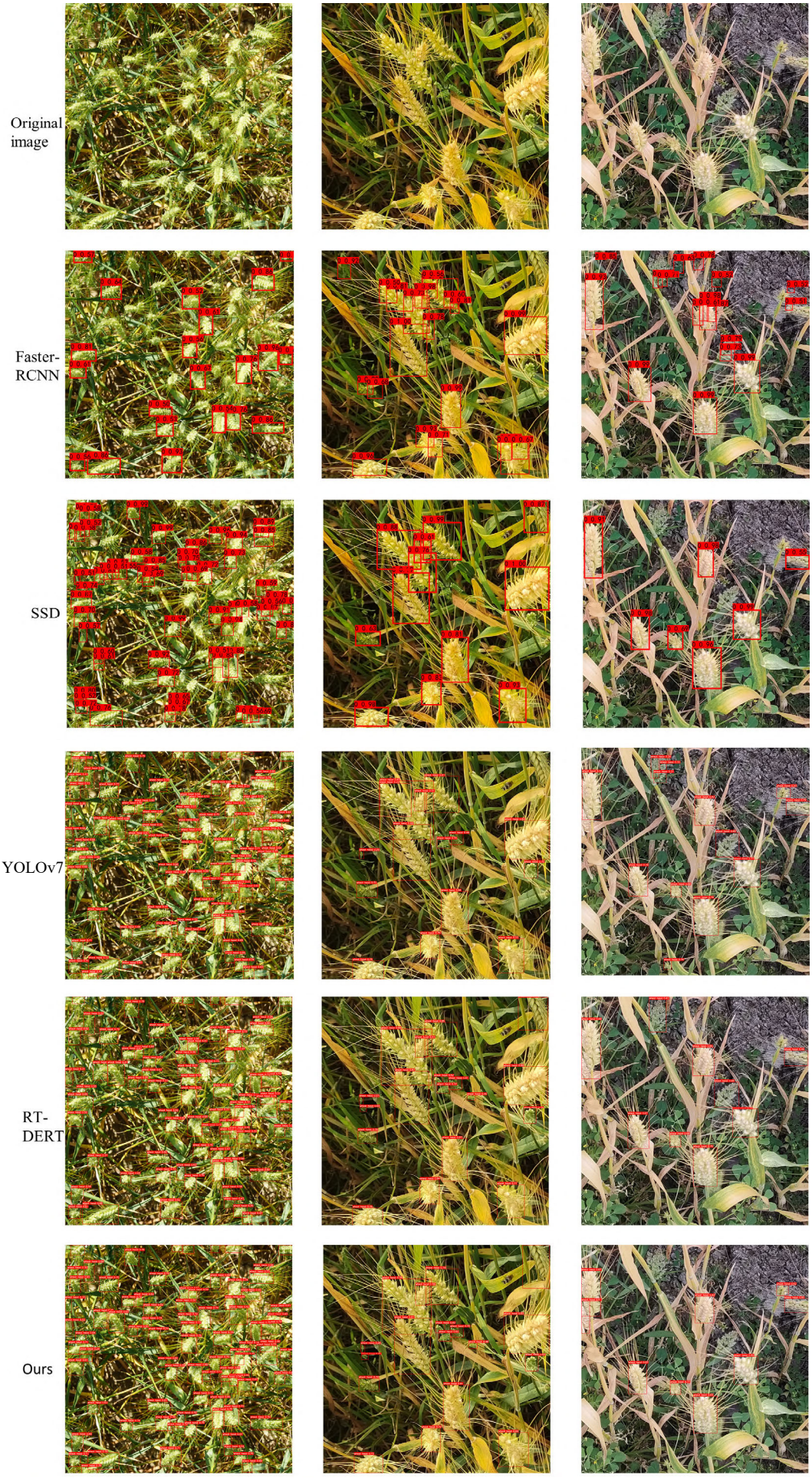


Fig 10 Visualization results of wheat spike detection

TABLE III
THE COMPARATIVE RESULTS OF MODEL TRAINING AFTER INTEGRATING DIFFERENT BACKBONE NETWORKS

Backbone	Precision	Recall	mAP50	Param	GFlops
ResNet18	91.1%	90.2%	93.4%	20.1M	58.08
RepViT	90.4%	88.0%	92.5%	13.3M	37.85
FasterNet	91.5%	90.5%	93.6%	16.8M	50.66
ShuffleNetV2	89.5%	87.0%	91.4%	9.3M	26.25
Ours	91.9%	91.1%	94.5%	22.2M	64.40

TABLE IV
COMPARISON OF MAINSTREAM OBJECT DETECTION MODELS

Models	Precision	Recall	mAP50	F1-score
Faster-RCNN	48.1%	47.5%	41.1%	47.8%
SSD	92.3%	43.2%	67.1%	58.8%
YOLOv7	88.9%	86.8%	90.8%	87.8%
RT-DETR	90.5%	90.2%	93.4%	90.3%
Ours	91.9%	91.1%	94.5%	91.5%

positives while minimizing false negatives. Additionally, the model achieves the highest F1-score of 91.5%, indicating a well-balanced trade-off between Precision and Recall.

Fig 10 illustrates the comparative detection performance of various models. Faster R-CNN and SSD demonstrate notable shortcomings, showing a high frequency of missed detections and false positives. In contrast, YOLOv7 and RT-DETR exhibit more consistent performance in densely populated areas of wheat spikes. However, their effectiveness diminishes in intricate or obstructed environments, leading to a continued presence of false alarms and missed detections. The proposed model shows improved accuracy in identifying wheat spikes, particularly in adverse environmental conditions. This enhanced performance is a result of design improvements that facilitate more effective feature extraction and differentiation between wheat spikes and surrounding elements. These findings validate the model's efficacy and its ability to perform well in practical agricultural settings.

Additional experiments were conducted to evaluate the robustness and generalization ability of the proposed method on images with adverse conditions, such as densely distributed wheat spikes and low illumination. The detection outcomes under these challenging scenarios are depicted in Fig 11 (Rectangles: missed detection; Diamonds: false detection; Circles: correct detection) YOLOv7 exhibits decreased performance under high-density conditions, with frequently missed detections in heavily overlapped areas of wheat spikes, indicating a limitation in distinguishing individual targets within densely clustered regions. In contrast, RT-DETR, while more sensitive in such conditions, experiences an increase in false positives, often misidentifying background textures as wheat spikes. In low-light environments, both YOLOv7 and RT-DETR show significant performance degradation, with elevated false positives and missed detections, suggesting compromised feature extraction capabilities when visual cues are less clear. The proposed method, however, maintains consistent and precise detection performance in both challenging scenarios due to enhanced feature representation and attention-guided localization mechanisms, enabling accurate differentiation of wheat spikes from complex backgrounds even in suboptimal lighting conditions. These findings highlight the superior adaptability and resilience of the proposed model, especially in agricultural settings where lighting and occlusion conditions can vary widely.

A series of experiments were conducted to validate the effectiveness of integrating object detection and multi-object tracking algorithms for wheat spike counting. The DeepSORT tracking algorithm was utilized to accurately track the motion trajectories of individual wheat spikes across consecutive video frames, assigning a unique identity (ID) to each spike. The experiments employed the enhanced RT-DETR model for real-time object detection to localize wheat spikes in each frame. Subsequently, DeepSORT was applied for precise tracking. The spike counting process utilized a multi-frame cross-line counting approach, as depicted in Fig 12.

To evaluate the counting performance of the proposed model, we conducted a comparative study against two established counting models, MCNN and TasselNetV2. An illustration of the evaluation images is presented in Fig 13. Evaluation metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). A lower MAE signifies a smaller average deviation from the ground truth, while a lower RMSE indicates reduced prediction variance. The R^2 metric assesses the alignment between predicted and actual values, with values closer to 1 denoting stronger consistency.

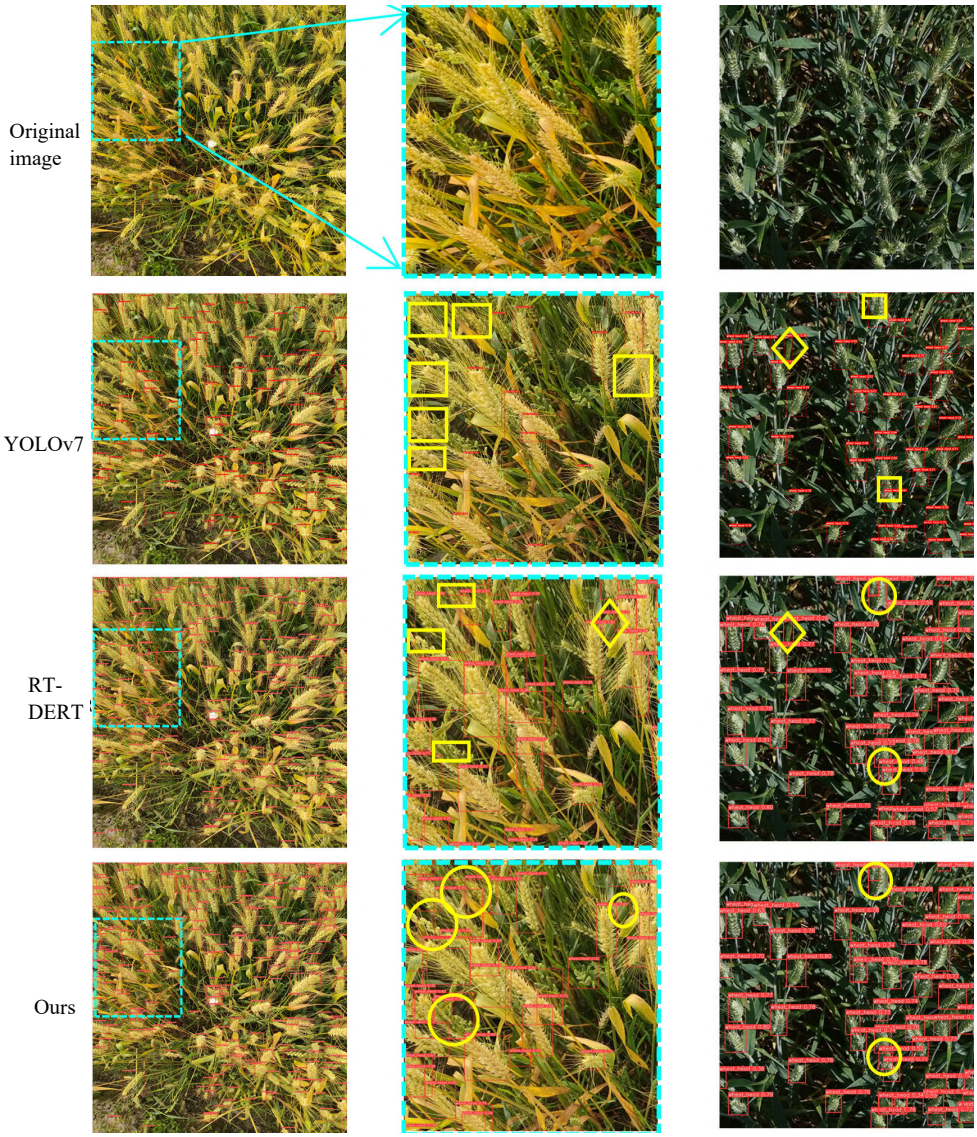


Fig 11 Visualization results of the challenging images

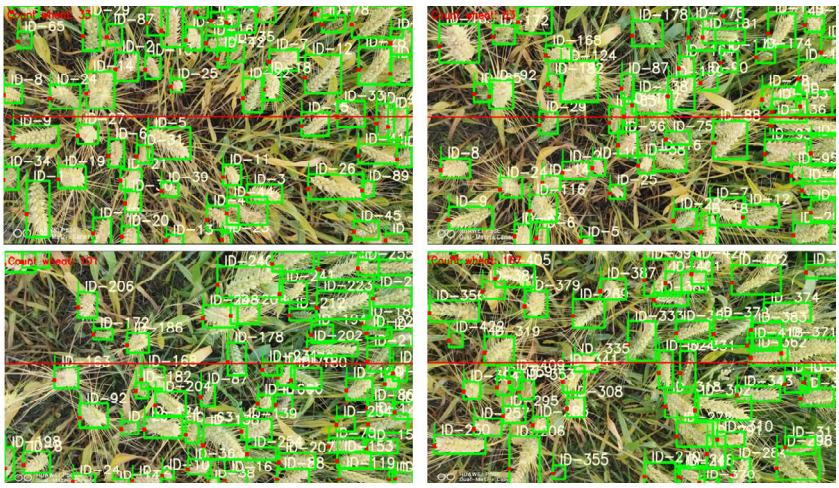


Fig 12 Visualization results of wheat spike tracking and counting

Fig 13 Original image and label image

All models underwent training and evaluation using the Global Wheat dataset, with results summarized in Table 5. The proposed model demonstrated an MAE of 9.20, an RMSE of 11.65, and an R^2 of 0.94, indicating high prediction accuracy and linear fit. These findings suggest that the proposed model outperforms MCNN and TasselNetV2 in counting performance[34][35][36].

TABLE V
COMPARISON OF THE PERFORMANCE OF DIFFERENT COUNTING MODELS

Model	MAE	RMSE	R^2
MCNN	11.09	14.32	0.90
TasselNetV2	9.72	13.38	0.92
Ours	9.20	11.65	0.94

V. CONCLUSIONS

In this investigation, the FasterNet, SENetV2, and RepGFPN components were integrated into the multi-object detection Transformer framework known as RT-DETR to tackle the intricate challenges associated with wheat spike detection. These challenges include the detection of small and densely clustered spikes, interference from weeds, and variations in color among wheat spikes, straw, and leaves at different growth stages. These enhancements were devised to enhance detection precision while minimizing computational burden. The integration of the FasterNet component effectively reduces computational complexity, thereby enabling real-time capabilities of the model. By incorporating the SENetV2 block within the FasterNet architecture, more precise extraction of both local and global features is achieved, enhancing the model's capacity to address partial occlusion and suppress interference from surrounding vegetation. Furthermore, the RepGFPN module boosts the network's performance by facilitating efficient extraction and fusion of multi-scale features, crucial for detecting wheat spikes of diverse sizes and shapes. Cumulatively, these adjustments yield a model with improved robustness and generalization, rendering it suitable for deployment in varied wheat cultivation settings and ensuring consistent performance under diverse field conditions.

While the current model has made significant advancements in wheat spike detection, there is a need to enhance its accuracy further. Future research will be guided by the design principles of the Adaptive Low-Pass Filter (ALPF) generator and Adaptive High-Pass Filter (AHPF) generator embedded in the Frequency-Aware Feature Fusion (FAFF) module [37]. Integrating ALPF into multi-scale feature fusion aims to address inconsistencies between low- and high-resolution features of the same category, thereby enhancing semantic coherence. Concurrently, AHPF is intended to improve the preservation of high-frequency detail information typically lost during down-sampling. Subsequent investigations will focus on combining ALPF and AHPF to optimize multi-scale feature fusion, reduce feature disparities, and enhance the extraction of intricate wheat spike details. Moreover, an expanded dataset comprising a larger number of field-acquired wheat spike images will be gathered to bolster the development of more resilient and broadly applicable detection methodologies.

REFERENCES

[1] OECD & FAO. (2023). OECD-FAO agricultural outlook 2023-2032.

[2] T. Liu, W. Wu, W. Chen, C. Sun, X. Zhu, and W. Guo, "Automated image-processing for counting seedlings in a wheat field," *Precision Agriculture*, vol. 17, no. 3, pp. 392-406, 2016.

[3] S. Hameed and I. Amin, "Detection of weed and wheat using image processing," in *Proc. 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pp. 1-5, 2018.

[4] J. X. Xue et al., "Wheat spike growth modeling based on a polygon," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 9, pp. 1175-1184, 2019.

[5] K. Bi et al., "Automatic acquisition characteristic parameters of wheat spike based on machine vision," in *Proc. 2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, pp. 148-154, 2011.

[6] J. A. Fernandez-Gallego et al., "Wheat spike counting in-field conditions: High throughput and low-cost approach using RGB images," *Plant Methods*, vol. 14, no. 1, pp. 1-12, 2018.

[7] F. Cointault and P. Gouton, "Texture or color analysis in agronomic images for wheat spike counting," in *Proc. 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pp. 696-701, 2007.

[8] Jiaqi Pan, Sutao Song, Yujie Guan, and Weikuan Jia, "Improved Wheat Detection Based on RT-DETR Model," IAENG International Journal of Computer Science, vol. 52, no. 3, pp705-719, 2025.

[9] S. Nema and A. Dixit, "Wheat leaf detection and prevention using support vector machine," in *Proc. 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pp. 1-5, 2018.

[10] Dhillon and G. K. Verma, "Convolutional neural network: A review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85-112, 2020.

[11] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," 2017 36th Chinese Control Conference (CCC), Dalian, China, pp. 11104-11109, 2017.

[12] Luyu Sun, and Yujun Zhang, "Steel Surface Defect Detection Based on YOLOv10," Engineering Letters, vol. 33, no. 5, pp1220-1231, 2025

[13] R. Li and Y. Wu, "Improved YOLOv5 wheat spike detection algorithm based on attention mechanism," *Electronics*, vol. 11, no. 11, p. 1673, 2022.

[14] S. Qing et al., "Improved YOLO-FastestV2 wheat spike detection model based on a multi-stage attention mechanism with a LightFPN detection head," *Frontiers in Plant Science*, vol. 15, p. 1411510, 2024.

[15] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE International Conference on Computer Vision*, pp. 1080–1088, 2015.

[16] M. A. Batin et al., "WheatSpikeNet: An improved wheat spike segmentation model for accurate estimation from field imaging," *Frontiers in Plant Science*, vol. 14, p. 1226190, 2023.

[17] L. Li et al., "Development of image-based wheat spike counter through a Faster R-CNN algorithm and application for genetic studies," *The Crop Journal*, vol. 10, no. 5, pp. 1303–1311, 2022.

[18] T. Wu et al., "Resspikech on the method of counting wheat spikes via video based on improved YOLOv7 and DeepSort," *Sensors*, vol. 23, no. 10, p. 4880, 2023.

[19] X. Li et al., "Traffic sign detection based on improved faster R-CNN for autonomous driving," *The Journal of Supercomputing*, vol. 78, no. 12, pp. 14145–14165, 2022.

[20] P. Sadeghi-Tehran et al., "DeepCount: In-field automatic quantification of wheat spikes using simple linear iterative clustering

and deep convolutional neural networks," *Frontiers in Plant Science*, vol. 10, p. 1176, 2019.

[21] D. Wang et al., "Combined use of FCN and Harris corner detection for counting wheat spikes in field conditions," *IEEE Access*, vol. 7, pp. 178930–178941, 2019.

[22] I. B. Parico and T. Ahamed, "Real time pspike fruit detection and counting using YOLOv4 models and Deep SORT," *Sensors*, vol. 21, no. 14, p. 4803, 2021.

[23] Y. Ge et al., "Tracking and counting of tomato at different growth period using an improving YOLO-DeepSort network for inspection robot," *Machines*, vol. 10, no. 7, p. 489, 2022.

[24] Z. P. Li et al., "Real-time detection and counting of wheat spikes based on improved YOLOv7," *Computers and Electronics in Agriculture*, vol. 218, p. 108670, 2024.

[25] Y. Zhao et al., "Detrs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

[26] M. Narayanan, "SENetV2: Aggregated dense layer for channelwise and global representations," *arXiv preprint* arXiv:2311.10807, 2023.

[27] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12021–12031, 2023.

[28] X. Xu et al., "Damo-YOLO: A report on real-time object detection design," *arXiv preprint* arXiv:2211.15444, 2022.

[29] E. David et al., "Global wheat head detection (GWHD) dataset: A large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods," *Plant Phenomics*, 2020.

[30] Z. Liu et al., "Faster-YOLO-AP: A lightweight apple detection algorithm based on improved YOLOv8 with a new efficient PDWConv in orchard," *Computers and Electronics in Agriculture*, vol. 223, p. 109118, 2024.

[31] L. Chen et al., "Frequency-aware feature fusion for dense image prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] A. Wang, H. Chen, Z. Lin, J. Han and G. Ding, "Rep ViT: Revisiting Mobile CNN From ViT Perspective," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 15909-15920,2024.

[33] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." Proceedings of the European conference on computer vision (ECCV). 2018.

[34] Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 589-597, 2016.

[35] Y. Li, X. Zhang and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 1091-1100, 2018.

[36] Xiong, Haipeng et al. “TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks.” Plant methods vol. 15 150. 11 Dec. 2019.

[37] L. Chen et al., "Frequency-aware feature fusion for dense image prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.