Improved Wheat Detection Based on RT-DETR Model

Jiaqi Pan, Sutao Song, Yujie Guan, Weikuan Jia

Abstract —**In agricultural production, accurately detecting wheat ears is critical for assessing crop yields and monitoring growth, among other key aspects. The traditional manual detection of wheat ears has the disadvantages of inefficiency, inaccuracy, and non-sustainability. This situation cannot adapt to the development needs of intensive and precise modern agricultural production. This work introduces an enhanced RT-DETR-based model for wheat ear recognition to improve accuracy in addressing these difficulties. The model incorporates a space-to-depth layer and a non-stride convolutional layer into the backbone network, improving its ability to capture subtle features in wheat ear images. This improvement enables the model to infer the complete shape of occluded wheat ears, thereby improving its ability to identify overlapping targets. Furthermore, the model replaces the convolution operations in the hybrid encoder with a Context Guided Block, introducing a context-guided mechanism that enhances feature learning and effectively distinguishes wheat ears from complex backgrounds. According to the experimental findings, the model has notable benefits in terms of detection accuracy. The optimized model obtains 93.5%, 54.5%, 91.2%, and 89.0% in terms of AP_{50}, AP_{50-95}, precision, and recall, respectively, according to evaluation results on the Global Wheat Dataset. This study can effectively meet the requirements of high precision and reliability of wheat spike detection in agricultural production, which offers robust assistance for intelligent planting monitoring and large-scale agricultural output management.**

Index Terms—**Wheat ear detection; RT-DETR model; space-to-depth layer; Context Guided Block**

I. Introduction

With the escalating threat to food security posed by global population growth and climate change, it has become particularly important to improve the accuracy of yield and yield forecasts for major food crops such as wheat [1]. The accurate recognition of wheat ears serves as a direct indicator for evaluating wheat yield, which is essential for informing agricultural production strategies and optimizing resource allocation [2]. Traditional methods for wheat ear detection predominantly rely on manual visual assessments, which are not only time-consuming and labor-intensive but also constrained by individual differences in expertise, thereby limiting the accuracy of the results and complicating the demands for large-scale and high-precision yield measurements [3]. Artificial Intelligence-based Visual Computing Technology Evolution presents a viable way to recognize images automatically [4]. However, there are substantial technological obstacles to efficient target detection and counting because of the high density distribution of wheat ears and the varied backgrounds inherent in real contexts [5].

Research on object detection techniques in densely populated scenes can provide valuable approaches for detecting wheat ears, especially considering the challenges posed by their dense distribution, occlusion, and overlapping during the growth process. He et al. [6] proposed the Hierarchical-NMS algorithm to address limitations of existing NMS and Soft-NMS algorithms in dense pedestrian detection, Effectively reducing the number of false positives and missing detections in dense situations. Improved non maximum suppression (NMS) can effectively solve the inter-target occlusion problem, however, this can easily lead to NMS incorrectly suppressing the candidate frames of different objects instead of their center frames when there is a dense concentration of targets with severe occlusion [7]. Wang et al. [8] utilized a novel bounding box regression loss that was specifically developed for crowded environments to achieve more robust localization. Although this exclusion loss function has shown effective results in pedestrian detection, it has been less frequently applied to dense agricultural scenarios, where many approaches rely on data augmentation for handling dense target detection. Du et al. [9] tackled the issue of dense clustering of small pests in field environments by employing cluster data generation, thereby extending the training dataset to enable the model to accurately identify regions with densely distributed target clusters. Data augmentation improves model robustness in dense detection by increasing sample diversity, but performing data augmentation operations requires additional computational resources and may limit the speed of the model [10]. Numerous scholars have responded to the challenges of dense detection by improving feature extraction and feature fusion techniques. Liu et al. [11] proposed an enhanced YOLOv5 model, which significantly improved the detection of densely distributed small fruits in citrus orchards through the introduction of a Coordinated

Manuscript received August 28, 2024; revised November 10, 2024.

This work is supported by Natural Science Foundation of Shandong Province in China (ZR2022MF349, ZR2020MF076); New Twentieth Items of Universities in Jinan (2021GXRC049).

J.Q. Pan is a postgraduate student of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (1660256685@qq.com);

S.T. Song is an associate professor of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (Corresponding author, e-mail: sutao.song@sdnu.edu.cn);

Y.J. Guan is a postgraduate student of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (1399973427@qq.com)

W.K. Jia is an associate professor of School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (Corresponding author, phone: +86-531-86181755; fax: +86-531-86181750; e-mail: jwk_1982@163.com)

Attention Module (CA [12]) and Bidirectional Feature Pyramid Network (BiFPN [13]). Additionally, Wen et al. [14] proposed the ST-YOLO model, integrating depthwise separable convolution with the Swin Transformer architecture to enhance target recognition in complex environments, demonstrating exceptional performance in agricultural applications such as tea bud detection and highlighting its potential for dense object detection scenarios. In dense and complex environments, the aforementioned research effectively enhanced the precision and resilience of dense object detection by modifying the NMS algorithm, data enhancement, and feature extraction technology. This solution resolved the issues of overlooked and false detections that are common in traditional methods.

The continuous advancement of dense detection technology has significantly influenced the specialized field of wheat ear detection, with numerous studies successfully applying innovative target detection algorithms that greatly enhance detection efficiency and accuracy. In early two-stage detection methods, Li et al. [15] explored the Faster R-CNN [16] for quantifying the density of spikes per unit area (SN) in wheat, demonstrating that this image recognition technique is applicable for genetic study on wheat SN. Although two-stage detection methods can yield high-precision results, they still exhibit computational redundancy in subsequent detection stages due to their inherent two-step strategy. Conversely, algorithms for one-stage identification, like those in the YOLO family, streamline the process by directly predicting object classes and locations within images, thereby significantly enhancing detection speed [17]. Wen et al. [18] introduced BiFPN, focus loss and attention module based on RetinaNet [19] to obtain a wheat spike detection and counting for SpikeRetinaNet, which achieved a mAP_{50} metric of 0.9262 on the Global Wheat Spike Detection (GWHD) dataset. The single-stage method has obvious advantages in inference speed, but in the localization and detection of small targets, it is prone to a high false detection rate due to the high number of densely generated candidate frames [15], while the anchorless frame detection network does not need to preset the anchor frame parameters in the detection and directly forecasts the location of the key point of the bounding box during the regression stage to further minimize the constraints of the predefined conditions [20]. Wang et al. [21] developed an anchor-free technique for detecting wheat ears with the attention-based ObjectBox [22], which significantly improves detection accuracy through enhanced feature connectivity, feature map fusion, and an optimized non-maximum suppression algorithm. The self-attention mechanism of Transformer [23] technology captures global dependencies, introducing new possibilities and substantial enhancements to wheat ear detection in complex environments. Zhou et al. [24] introduced a wheat ear detection network employing Transformer architecture, integrating multi-window features and a feature pyramid network for extracting multi-scale features and implementing efficient self-attention, thereby enhancing the efficacy of wheat ear recognition under intricate field circumstances.

While existing models demonstrate high accuracy in wheat target detection, recognizing wheat ears at varying scales remains challenging, particularly when dealing with small targets and limited data availability for these small targets. This study introduces an enhanced method for detecting wheat ears with the Real-Time Detection Transformer (RT-DETR [25]) model. RT-DETR retains the strengths of the DETR model [26] in handling complex image scenes, making it better suited for the demands of agricultural monitoring. The main contributions of this study include:

(1) To mitigate the information loss resulting from the stride-2 depthwise separable convolution during feature extraction, a space-to-depth layer and a non-stride convolutional layer are incorporated to modify the spatial dimensions of feature maps. This approach maintains high spatial resolution while improving processing efficiency and accuracy, enabling better handling of overlapping region details.

(2) To tackle the challenges presented by complex backgrounds in wheat field detection, the convolution operations within the hybrid encoder are replaced with Context Guided Blocks (CGBlocks). This adjustment enhances the ability to learn contextual information, thereby improving effectiveness in managing background complexity during wheat ear detection.

(3) Even with a limited amount of data for small-scale targets, the model can still achieve high detection accuracy, indicating that it possesses strong generalization capabilities and robustness, thereby enabling effective detection of wheat ears in complex environments.

II. Methods

The growing conditions and environment of wheat ears make target detection difficult. Accurate detection of wheat ears can furnish dependable data support for agricultural production, allowing agriculturists and administrators to make more informed decisions based on actual crop growth conditions. This study enhanced the RT-DETR model to augment the accuracy of wheat target identification. The RT-DETR model is an end-to-end real-time object detection model based on the Transformer architecture. The optimized model still has the same architecture as the original model: backbone network, hybrid encoder, and Transformer decoder with auxiliary prediction head. First, the space-to-depth layer (SPD-Conv [27]) module is incorporated into the backbone network to augment the ability to capture intricate details in wheat ear images, hence enhancing the recognition of overlapping targets. In addition, by replacing the convolution operation in the hybrid encoder with the Context Guided Block(CG Block [28]), a context guided mechanism is implemented to further augment the effect of feature learning and effectively distinguish between wheat ear targets and complex backgrounds. Figure 1 exhibits the architecture of the wheat detection algorithm with these adjustments, demonstrating the optimization of its structure and enhancement of performance.

A. Feature Extraction Network

In wheat ear detection, the feature extraction network is essential for automatically identifying and extracting important characteristics of wheat ears from the input

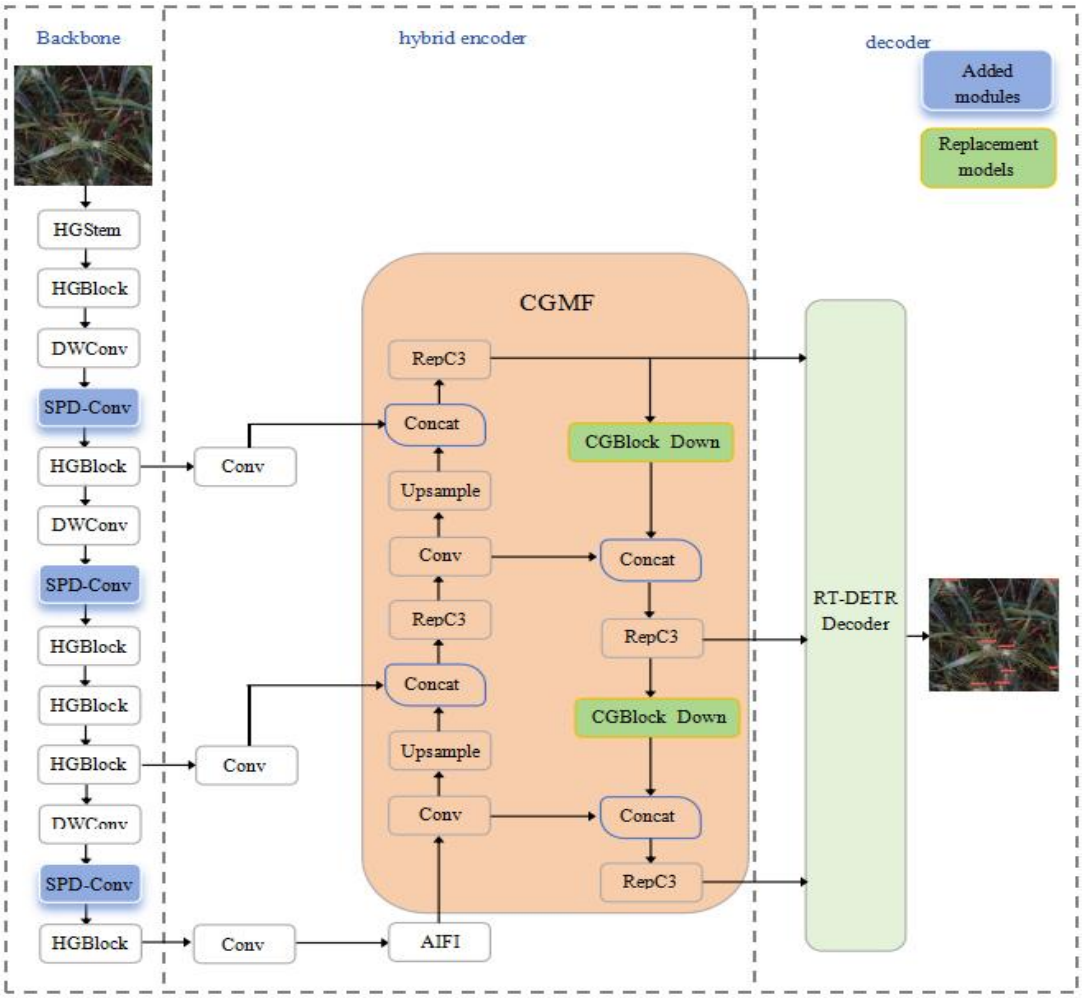


Fig.1 Framework diagram of enhanced wheat ear detection model using RT-DETR

images, including shape, size, color, and texture. These features are essential for subsequent classification, localization, and analysis, particularly in automated and precision agriculture, where accurate detection of wheat ears can significantly enhance the efficiency of crop evaluation and processing.

1) Backbone Network

The backbone of RT-DETR uses HGNet-v2 for feature extraction, and the overall structure consists of a Stemmodule and multiple HGblock modules. The backbone network performs detailed feature processing and gradual refinement of the image through four main stages. The first stage of the backbone starts with the input layer HGStem, which performs preliminary feature extraction on the original wheat ear image and increases the channel count from 32 to 48. Afterwards, the channel count was augmented to 128 by HGBlock to enhance image detail acquisition. The second stage uses Depthwise Separable Convolution (DWConv [29]) for spatial downsampling and expands the features from 96 to 512 channels through HGBlock, thereby increasing the details and complexity of the feature map. The third stage further employs DWConv to downsample to a deeper 1024 channels, followed by complex feature integration and enhancement through three HGBlocks with different configurations to ensure effective learning and extraction of high-level features. In the final stage, DWConv is used for subsampling again, and finally, a large-capacity HGBlock is used to upgrade the features to 2048 channels to capture the most complex environmental information and object semantic information, providing support for high-precision wheat image understanding and detection.

In the real environment where wheat is cultivated, the dense growth that occurs among the ears in the wheat field can easily cause the detection algorithm to mistake the features of neighboring ears together, which in turn triggers the problems of misdetection and omission. Within the original framework of the RT-DETR model, the backbone network utilizes DWConv with a stride of 2 to shrink the spatial size of the feature map, all while concurrently enhancing the feature depth. This approach may lead to the omission of essential spatial details, especially when processing intricate images of wheat ears requiring high-resolution accuracy.

2) SPD Module

To mitigate the information loss induced by DWConv, the SPD-Conv module is implemented to modify the spatial dimensions of the feature map. By setting the stride of DWConv is altered to 1, thereby preventing the spatial size of the feature map from diminishing through convolution.

SPD-Conv initially executes a space-to-depth convolution process, partitioning the feature map X into several sub-feature maps$\{f_{i,j}\}$. The cutting approach involves acquiring several sub-feature maps by sampling the main feature map at predetermined intervals based on the designated scale map. Each sub-feature map is created by downsampling the original feature map. Eq.1 is the

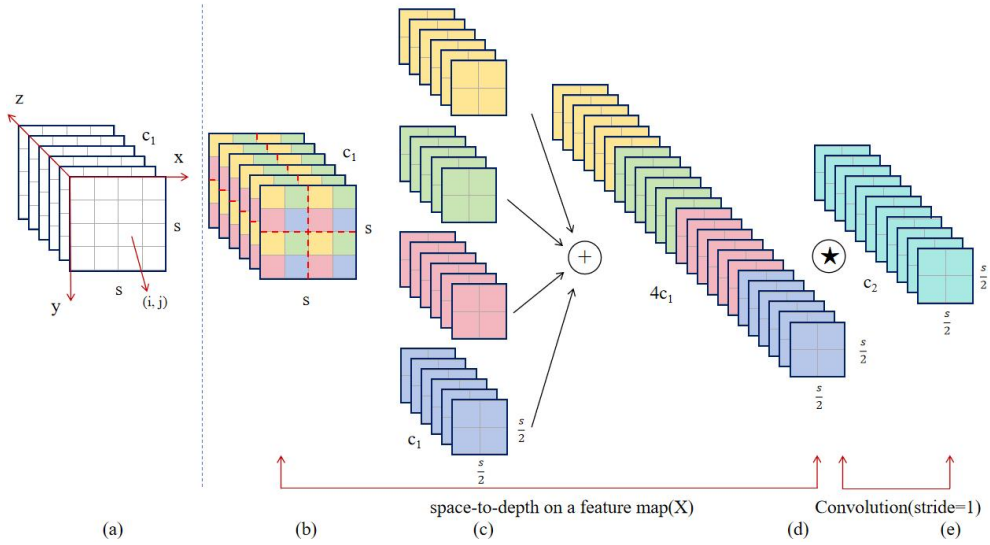


Fig. 2. SPD-Conv structure diagram when Scale is 2

sub-feature sequence cut out at the scale:

$$\begin{cases} f_{0,0} = X[0{:}\,S{:}\,scale,\ 0{:}\,S{:}\,scale]\,, \\ f_{1,0} = X[1{:}\,S{:}\,scale\,,\ 0{:}\,S{:}\,scale], \\ \cdots, \\ f_{scale-1,0} = X[scale-1{:}\,S{:}\,scale,\ 0{:}\,S{:}\,scale]; \end{cases} \tag{1}$$

Where Eq. 1 indicates that the original feature map X starts from the upper left corner (i.e. the starting index is 0,0) and selects it every S pixels until the multiple of scale is reached to form the first sub-feature graph $f_{0,0}$. And so on, until $f_{scale-1,0}$, the feature map X starts at the position where it is offset by scale-1 pixels in the row direction, forming the last sub-feature map $f_{scale-1,0}$.The same process is applied to the column direction as well, and Eq. 2 denotes denotes the change in the feature map along the column direction:

$$\begin{cases} f_{0,1} = X[0{:}\,S{:}\,scale,\ 1{:}\,S{:}\,scale], \\ f_{1,1} = X[1{:}\,S{:}\,scale,\ 1{:}\,S{:}\,scale], \\ \cdots, \\ f_{scale-1,1} = X[scale-1{:}\,S{:}\,scale,\ 1{:}\,S{:}\,scale]; \\ \vdots \\ f_{0,scale-1} = X[0{:}\,S{:}\,scale\,,\ scale-1{:}\,S{:}\,scale], \\ f_{1,1-scale} = X[1{:}\,S{:}\,scale,\ 1-scale{:}\,S{:}\,scale]\,, \\ \cdots, \\ f_{scale-1,scale-1} = X[scale-1{:}\,S{:}\,scale,\ scale-1{:}\,S{:}\,scale]; \end{cases} \tag{2}$$

$f_{0,1}$ represents the production process of the next subfeature map that begins after 1 pixel shift in the column direction. And so on until $f_{scale-1,scale-1}$, which is where the original feature map X begins with a scale-1 pixel shift in both row and column directions, forming the final subfeature map.

A non-strided (i.e., stride 1) convolutional layer operation is incorporated following the SPD feature transformation layer. This convolutional layer has C_2 filters, where $C_2 < scale^2 C_1$, and further transforms the intermediate feature map $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1)$ into $X''(\frac{S}{scale}, \frac{S}{scale}, C_2)$. Non-strided convolution is used in order to preserve as much discriminative feature information as possible.

Figure 2 shows the process of processing an intermediate feature map X of any size using interval sampling with a scale of 2, resulting in four sub-feature maps $f_{0,0}$, $f_{1,0}$, $f_{0,1}$ and $f_{1,1}$, each of which has the shape of $(\frac{S}{2}, \frac{S}{2}, C_1)$. To create a new feature map $X'(\frac{S}{2}, \frac{S}{2}, 4C_1)$, all sub-feature maps are concatenated along the channel axis. Then $X''(\frac{S}{2}, \frac{S}{2}, C_2)$ is obtained by doing a non-strided convolution using C_2 filters.

In dealing with dense wheat ear detection scenarios, especially when the targets are small and closely arranged, it becomes crucial to maintain high spatial resolution information. This is because excessive downsampling may cause the details of small objects to be blurred, making these objects difficult to be accurately identified and localized in deep networks. The integration of the SPD-Conv module into the backbone network not only improves the recognition of small objects but also ensures the maintenance of high spatial resolution while enhancing processing efficiency and accuracy, further meeting the detection requirements in complex environments.

B. Feature Fusion Network

The wheat spike detection network optimized based on RT-DETR model handles multi-scale features through hybrid encoders at the neck, especially in processing wheat spike targets at different scales, and subsequently realizes feature fusion among different scales to improve the performance of wheat spike detection. The hybrid encoder in the neck comprises two components: the Attention-based Intra-scale Feature Interaction (AIFI [25]) module and the CNN-based Cross-scale Guided Feature-fusion Module (CGFM).

1) Feature Fusion

The hybrid encoder receives the extracted high-level features of wheat ears from the feature extraction network, commencing with the alteration of the channel numbers in the feature map through 1×1 convolution, and subsequently employing AIFI and Repeated 3×3 Convolution (RepC3) to augment feature representation.The AIFI module performs self-attention operations on high-level features and processes among features within the same scale, thereby helping subsequent modules to better detect and recognize wheat ears in images. The convolutional layer in the CNN-based Cross-scale Feature-fusion Module (CCFM

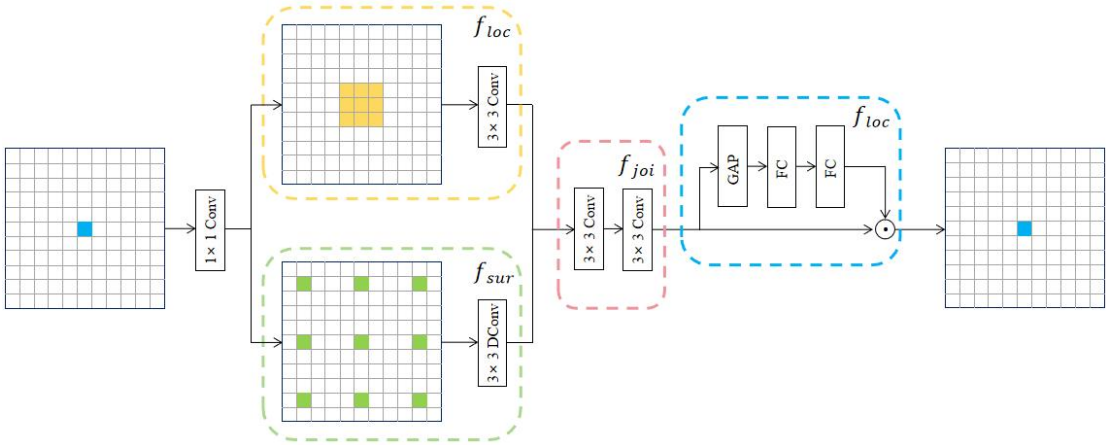


Fig. 3. CG Block structure diagram

[25]) in the neck hybrid encoder may not be able to fully capture and utilize local and global contextual information when processing images. In the intricate and dynamic context of the wheat field, precisely distinguishing the wheat ears from other background elements proves challenging. Consequently, GC Block is incorporated into the CCMF module to create the CGFM module.

In CGMF, the amalgamation of up-sampling and down-sampling processes with element-wise concatenation (Concat) enables multi-scale feature fusion. This enables the model to effectively consolidate information from various regions and hierarchical levels of detail within the image. The hybrid encoder with CG Block enhances the detection capability of at diverse scales and improves the information fusion efficiency of low-level and high-level characteristics via route aggregation, enabling the model to more effectively identify and locate wheat ears in images.

2) Context Guided Block Module

To reliably recognize wheat ears in intricate backgrounds amid varying lighting conditions, the model can comprehend and incorporate this contextual information to boost the precision and resilience of detection. This paper introduces a context-guided mechanism by replacing the Conv operation in the CCFM module with a Context Guided Block and names this module CGMF to address the background complexity problem in the wheat ear dataset.

CG Block consists of local feature extractor $f_{loc}(*)$, surrounding context extractor $f_{sur}(*)$, joint feature extractor $f_{joi}(*)$, and global context extractor $f_{glo}(*)$. The structural framework is shown in Figure 3. First, the initial feature map X undergoes an initial convolution process, followed by batch normalization and the application of a PReLU activation function for feature conversion and downsampling, which can be expressed as:

$$X_{down} = PReLU(BN(Conv(X))) \quad (3)$$

Where, $Conv$ denotes the convolution process, using a 3×3 convolution kernel with a stride of 2. Subsequently, local feature extraction is executed, and channel-wise convolution is applied to the subsampled feature map X_{down} to derive local feature F_{loc}. The above operation formula is as follows:

$$F_{loc} = C_{loc}(X_{down}) \quad (4)$$

In parallel, the surrounding context feature extraction F_{sur} is performed, and X_{down} is also processed by channel-wise convolution with dilation rate to control the size of the receptive field for capturing surrounding context features:

$$F_{sur} = C_{sur}(X_{down}) \quad (5)$$

Then perform joint feature extractor F_{joi} to fuse local feature extractor F_{loc} and surrounding context extractor f_{sur} in the channel dimension to guarantee that the model understands the data of each pixel or local area, and also understand the relationship between these regions in the overall context, and through batch normalization and PReLU activation function:

$$F_{joi} = PReLU(BN(F_{loc} \oplus F_{sur})) \quad (6)$$

Where $\oplus$ represents feature fusion (channel splicing). The integrated features are dimensionalized using a 1×1 convolution layer to minimize the parameter count and computational complexity:

$$F_{red} = C_{reduce}(F_{joi}) \quad (7)$$

Finally, global feature modulation is carried out in order to amplify significant features and reduce the significance of less significant features:

$$F_{out} = F_{red} \cdot \sigma(FC(AvgPool(F_{red}))) \quad (8)$$

Where, σ denotes the Sigmoid function, utilized to produce the important weight for each channel in the feature graph, AvgPool represents the adaptive average pooling procedure, whereas FC signifies the fully connected layer.

The f_{loc} and f_{sur} extractors employ channel convolution to cut computational expenses across channels without compromising accuracy. The Conv operation in the hybrid encoder of the RT-DETR model is substituted with the CG Block, which facilitates successful fusing and downsampling of the input feature map while enhancing the efficiency of the model in leveraging local and global information. This helps the model to enhance its comprehension of the semantic information and contextual linkages within the image, so more efficiently differentiating the connections between the wheat ear target and the intricate background. Give the model greater insight into the contextual linkages and semantic content of images so that it can discriminate between complicated backdrops and wheat ear targets more successfully. By reducing the amount of network parameters, this replacement can significantly

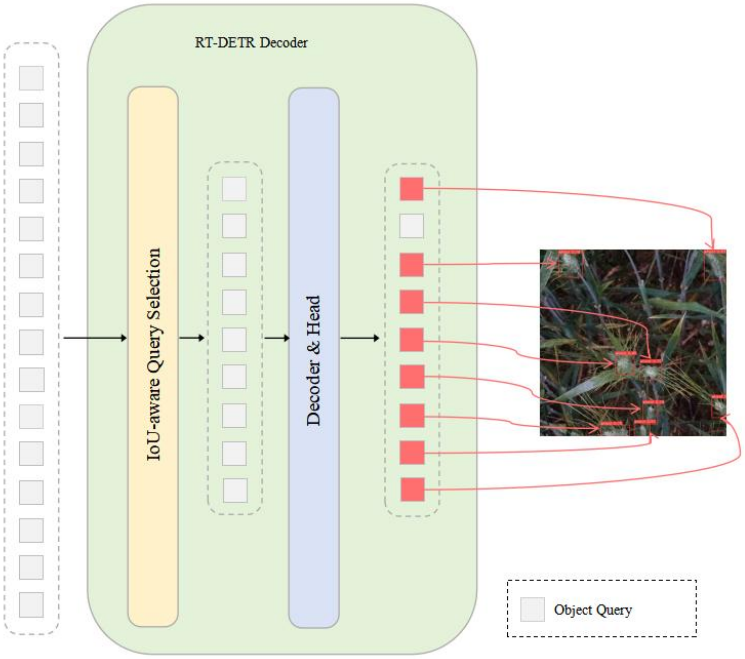


Fig.4. Decoder structure diagram

increase the detection performance and generalization capacity while also increasing its operational efficiency.

C. Decoder

The decoder section of the wheat ear detection network optimized based on the RT-DETR model adopts the standard Transformer Decoder structure, consisting of multiple Transformer decoding layers, which are used to process input features and generate the final prediction results. By stacking multiple decoding layers, the model can gradually extract and organize features to achieve more accurate wheat ear detection.

Employ IoU-aware Query Selection to select a predetermined quantity of image characteristics as the initial object query for the decoder. It is important to take into account both the features with elevated categorization scores and their IoU scores when doing object queries. The model chooses features with elevated classification and IoU metrics to initialize decoder queries. This can reduce the selection of prediction boxes that have high classification scores but poor alignment with the real bounding boxes. Lastly, the object query is repeatedly optimized by the decoder with auxiliary prediction head to produce boxes and confidence scores. Figure 4 illustrates the decoder framework diagram:

D. Loss

The loss function in the RT-DETR model is composed of multiple components, including classification loss(L_{cls}) , bounding box regression loss (L_{box}) and Generalized Intersection over Union Loss [30] (L_{GIoU}). The overall loss function of the model is presented as Eq. 9:

$$L(\hat{y}, y) = \lambda_1 L_{cls} + \lambda_2 L_{box} + \lambda_3 L_{GIoU} \quad (9)$$

where λ_1 =1, λ_2 =5 and λ_3 =2, denoting the weights of classification loss, bounding box localization loss, and bounding box overlap loss. By adjusting these weights, it is possible to influence the focus of the model during the learning process. Specifically, when λ_1=1, it indicates that the weight of the classification loss is relatively small, primarily because the targets in this scenario are relatively homogeneous. If the categorization loss weights are too high, the model may focus too much on the correctness of the category and ignore the accuracy of the location. Therefore, the weight of the categorization loss is set to a smaller value to balance it with the other losses. The value of λ_2=5reflects a greater emphasis on the precision of bounding box localization during the optimization process. This is because the cost of errors in object location prediction is typically higher, especially in high-precision object detection tasks. Therefore, assigning a higher weight to the bounding box loss encourages the model to more effectively learn the accurate prediction of object locations. The value of λ_3 =2 is slightly higher than the classification loss but lower than the bounding box loss. This is because GIoU acts mainly on the optimization of the overlap of bounding boxes rather than the direct bounding box position or size. Therefore, while it helps to improve positioning accuracy, its role needs to be synergistic with the bounding box loss and not overemphasized, as this can cause the model to deviate from the correct adjustment of the bounding box position during the optimization process.

In the wheat ear target detection model, since the target detection category is only wheat ears, the classification task is a binary classification problem, that is, detecting wheat ears and background. The wheat target detection model may successfully reduce background interference and increase wheat target recognition accuracy by employing Focal Loss, particularly in scenes with a complex background and few wheat ears.The loss function is shown in Eq. 10:

$$L_{cls} = Focal\ Loss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (10)$$

Where p_t is the probability that the model predicts the correct category, α_t is a balancing factor that controls the proportion of positive and negative samples to cope with category imbalance, and γ is a focusing factor that adjusts

(a) Similar colors (b) Overlapping targets (c) complicated background

(c) Backlight (d) Frontlight (e) Shadow occlusion

Fig.5. Characteristics of global wheat ear data

TABLE I
THE NUMBER OF WHEAT EARS OF DIFFERENT SIZES IN THE WHEAT EAR RECOGNITION DATA SET

Dataset	Class	Images	Instances	Target Amount		
				Small ($0<area<32^2$)	Medium ($32^2<area<96^2$)	Large ($96^2<area$)
Train Set	wheat	2,799	129,856	1,568（1.20%）	106,541(82.05%)	21,747(16.75%)
Test Set	wheat	700	32,298	491(1.52%)	26,685(82.62%)	5,122(15.86%)

the loss contribution of difficult and easy samples.

The bounding box regression loss quantifies the coordinate discrepancy between the predicted box and the actual box, employing the Mean Absolute Error (MAE) loss, which is determined by averaging the absolute deviations between the predicted and real values, thereby assessing the divergence between the predictions of the model and the true outcomes. The loss function is represented in Eq. 11:

$$L_{box} = \frac{1}{N}\sum_{i=1}^{N} |\widehat{b_i} - b_i| \tag{11}$$

Where $\widehat{b_i}$ represents the anticipated bounding box coordinates, while b_i denotes the actual bounding box coordinates of the *i*-th target.

The Generalized Intersection over Union Loss (GIoU Loss) enhances bounding box regression by incorporating the notion of enclosing boxes, hence overcoming the deficiency of IoU in generating gradients when the predicted box lacks overlap with the ground truth box. The loss function is shown in Eq. 12:

$$L_{GIoU} = 1 - \frac{Area(\widehat{b_i} \cap b_i)}{Area(\widehat{b_i} \cup b_i)} + \frac{Area(\mathrm{C}) - Area(\widehat{b_i} \cup b_i)}{Area(\mathrm{C})} \tag{12}$$

Where C denotes the minimal enclosing rectangle that contains both the expected and actual bounding boxes, $Area\ (\widehat{b_i} \cap b_i)$ indicates the region of overlap between the anticipated and real bounding boxes, and $Area\ (\widehat{b_i} \cup b_i)$ is the area of their union.

The structure of the loss function allows for simultaneous optimization of bounding box localization and object classification during training. By integrating scores, it strengthens the consistency between classification and localization for positive samples, thereby enhancing overall detection performance and facilitating a deeper understanding and learning of the content within images.

III. RESULTS AND ANALYSIS

A. Dataset

The dataset was derived from the Global Wheat Head Dataset (GWHD2021) [31] and consisted of 3,499 RGB images (1024×1024 pixels), totaling 162,154 ears. These wheat images originate from various areas. Owing to varying shooting angles and lighting conditions,

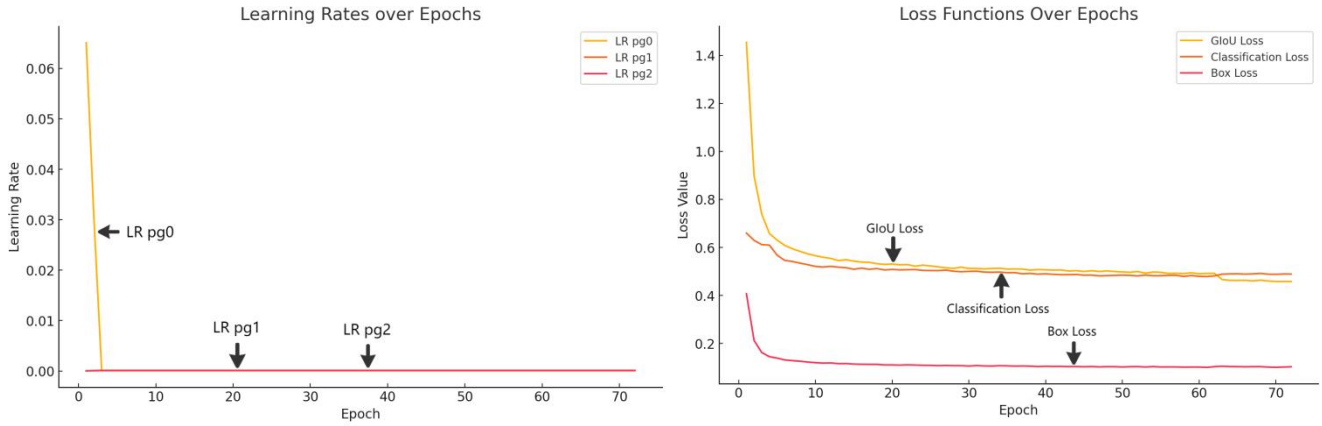


Fig.6. Left: Learning Rate Value Curve.
Right: Loss Value Curve.

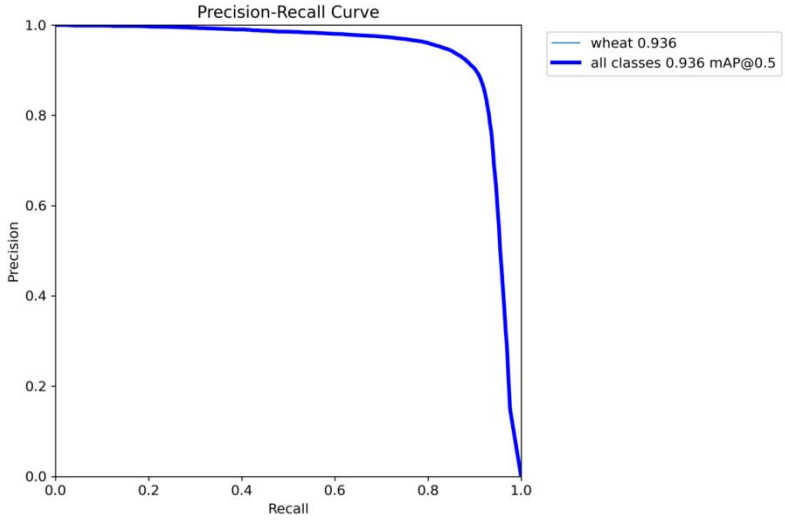


Fig.7. P-R curve of model training

there exists variation and complexity in the datasets due to varying wheat growth periods, distribution densities, and ear sizes. The GWHD2021 dataset is split at random into a training subset of 2,799 images and a validation subset consisting of 700 images, maintaining an 8:2 ratio. A portion of the wheat image is depicted in Figure 5.

According to the definition of coco data set, The target area smaller than 32×32 pixels is designated as a small target; the target area ranging from 32×32 to 96×96 pixels is classified as a medium target; the target area exceeding 96×96 pixels is referred to as a large target. Additionally, the quantity of wheat ears of varying sizes in the wheat ear recognition dataset is enumerated. In the training set of 2,799 photographs, there are a total of 129,856 wheat ear targets based on their proportion in the images. A total of 1,568 (1.20%) wheat ears were identified as small targets, 106541 (82.05%) as medium targets, and 21,747 (16.75%) as giant targets; In the test set of 700 photos, there are a total of 32,298 targets, with 491 (1.52%) wheat ears identified as small targets, 26,685 (82.62%) wheat ears identified as medium targets, and 5,122 (15.86%) wheat ears identified as large targets.

B. Experimental setup and model optimization strategies

The experiments involved in this paper are mainly configured with Ubuntu 16.04 LTS operating system and Intel® Xeon(R) Silver 4214R CPU @ 2.40GHz × 45. Running on 10GB GPU NVIDIA GeForce RTX 3080 and V11.4 CUDA, all models use Python version 3.8 and Pytorch version 1.12. The model components were constructed by MMdetection v2.2.0 learning library.

In this experiment, all images were resized to 640×640 for input to the network to standardize the criteria. This research establishes an initial learning rate of 0.0001 to optimize network parameters and minimize the loss function, utilizing Adaptive Moment Estimation with Weight-decay (AdamW) to facilitate the training process. The adaptive learning rate mechanism and weight decay strategy of AdamW contribute to enhancing training stability and preventing overfitting, while effectively accommodating the optimization needs of complex network structures. The graph depicting learning rate fluctuations is presented in Figure 6. Left, where the three parameter groups (pg0, pg1, and pg2) follow different learning rate adjustment strategies. The most obvious feature is reflected in pg0, which has a high initial learning rate and decays rapidly to near zero in early iterations. In contrast, the learning rates of pg1 and pg2 remained relatively stable in the initial stages and tended to remain constant and low after the decline. This learning rate strategy facilitates a balance between convergence velocity and final model correctness, particularly in intricate training tasks, and effectively mitigates issues such as gradient vanishing or explosion.

TABLE II
ABLATION EXPERIMENT RESULTS

Model	SPD-Conv	CG Bolock	Precision	Recall	AP_{50}	Ap_{50-95}
RT-DETR	—	—	90.5	87.5	91.9	50.7
	√	—	90.7	88.9	93	53.5
	—	√	90.7	90.5	92.4	53.4
	√	√	91.2	89.0	93.5	54.5

Combined with the characteristics of AdamW optimizer, the model is trained after 72 epochs, and set the batch size to 4, the momentum parameter to 0.9, and the weight decay coefficient to 0.0001. These optimization methodologies afford the model sufficient time and chance to assimilate the data features, enabling it to attain high detection accuracy during a reduced training epoch and exhibit robust generalization capability, so successfully mitigating the risk of underfitting. Figure 6 Right illustrates the training loss curve of the improved RT-DETR model on the dataset, utilizing the specified parameters. The rapid decrease in GIoU Loss and Box Loss reflects a significant improvement in the ability to predict bounding boxes, while the relatively smoother decline in classification loss highlights the inherent complexity of the category classification task. As training proceeds, the loss functions converge gradually and the overall performance of the model stabilizes.

C. Evaluation metrics

This research use precision (P) and recall (R) as criteria for assessing model performance. The corresponding formulas are presented in Eq. 13 and Eq. 14:

$$P = \frac{TP}{TP + FP} \cdot 100\% \quad (13)$$

$$R = \frac{TP}{TP + FN} \cdot 100\% \quad (14)$$

In Eq. 15, TP denotes the quantity of actual wheat ears accurately identified as such by the model, signifying the true positive instances where the wheat ear targets are correctly recognized by the model. FP denotes the quantity of background or non-wheat areas erroneously classified as wheat ears by the model. FN denotes the quantity of actual wheat ears that the model could not identify, signifying the overlooked wheat ear targets. Figure 7 illustrates the Precision-Recall (P-R) curve throughout the model training procedure.

$$AP = \int_0^1 P(r)dr \quad (15)$$

For the GWHD2021 dataset, the model evaluates only one category, so the AP itself represents the overall performance of the model on that particular task, with the same concept as calculating the mean of multiple categories (mAP), where the mean accuracy and mean average precision are calculated as follows.

This study establishes an IoU threshold range from 0.5 to 0.95, utilizing a step size of 0.05, resulting in a total of 10 thresholds. The mean value of these thresholds is used as the evaluation metric, Average Precision (AP). Simultaneously, 101 recall points are selected within the range of [0, 0.01, ..., 1], and the corresponding precision values are averaged to obtain the average precision under the 0.5 and 0.75 thresholds. In addition to introducing the two metrics AP_{50} and AP_{75}, the average accuracies of the small-scale, mesoscale, and large-scale targets are defined as AP_s, AP_m and AP_l, respectively, to better illustrate the detection performance of wheat ears at various sizes; Accordingly, the Average Recall for small-scale, mesoscale, and large-scale objects is defined as AR_s, AR_m, and AR_l, respectively. Finally, the number of parameters is also employed as a gauge to thoroughly assess the complexity of the model.

D. Ablation experiment

This study conducts ablation experiments to rigorously assess the effectiveness of the SPD-Conv and CG Block modules. The SPD-Conv module is initially incorporated into the baseline model, followed by a comparative examination of detection outcomes on the worldwide wheat dataset, both prior to and subsequent to its deployment. Subsequently, the CG Block module is added for further comparison.

A stepwise introduction strategy is adopted to explore the differential impacts of various modules on object detection performance in depth. The baseline model RT-DETR attains an accuracy of 90.5%, a recall rate of 87.5%, an AP_{50} of 91.9%, and an AP_{50-95} of 50.7%, as presented in Table 2. Upon incorporating the SPD-Conv module into the backbone network, the accuracy, recall rate, AP_{50}, and AP_{50-95} improved by 0.2%, 1.4%, 1.1%, and 2.8%, respectively, in comparison to the previous version. The precision and AP_{50} were significantly enhanced to 90.7% and 93%, respectively, indicating an overall improvement in performance, particularly in the AP_{50} and AP_{50-95} metrics, which suggests that this module has strengthened detection capability. When the convolution operation in the neck of the network is replaced with the CG Block, the accuracy, recall rate, AP_{50}, and AP_{50-95} improve by 0.2%, 3.0%, 0.5%, and 2.7%, respectively, compared to the previous model. Although the recall rate increases to 90.5%, the AP_{50} experiences a slight decrease to 92.4%. When both the SPD-Conv and CG Block modules are integrated simultaneously with the baseline model, the performance metrics reach their optimal levels: an accuracy of 91.2%, a recall rate of 89.0%, an AP_{50} of 93.5%, and an AP_{50-95} of 54.5%. These results indicate that the effective synergy between these two modules significantly enhances object detection performance.

This report presents the findings of the ablation experiment to allow for a more direct comparison of the influence of each module on the detection outcomes. The visualization results are presented in Figure 8. The yellow highlighted box denotes the misidentified region,

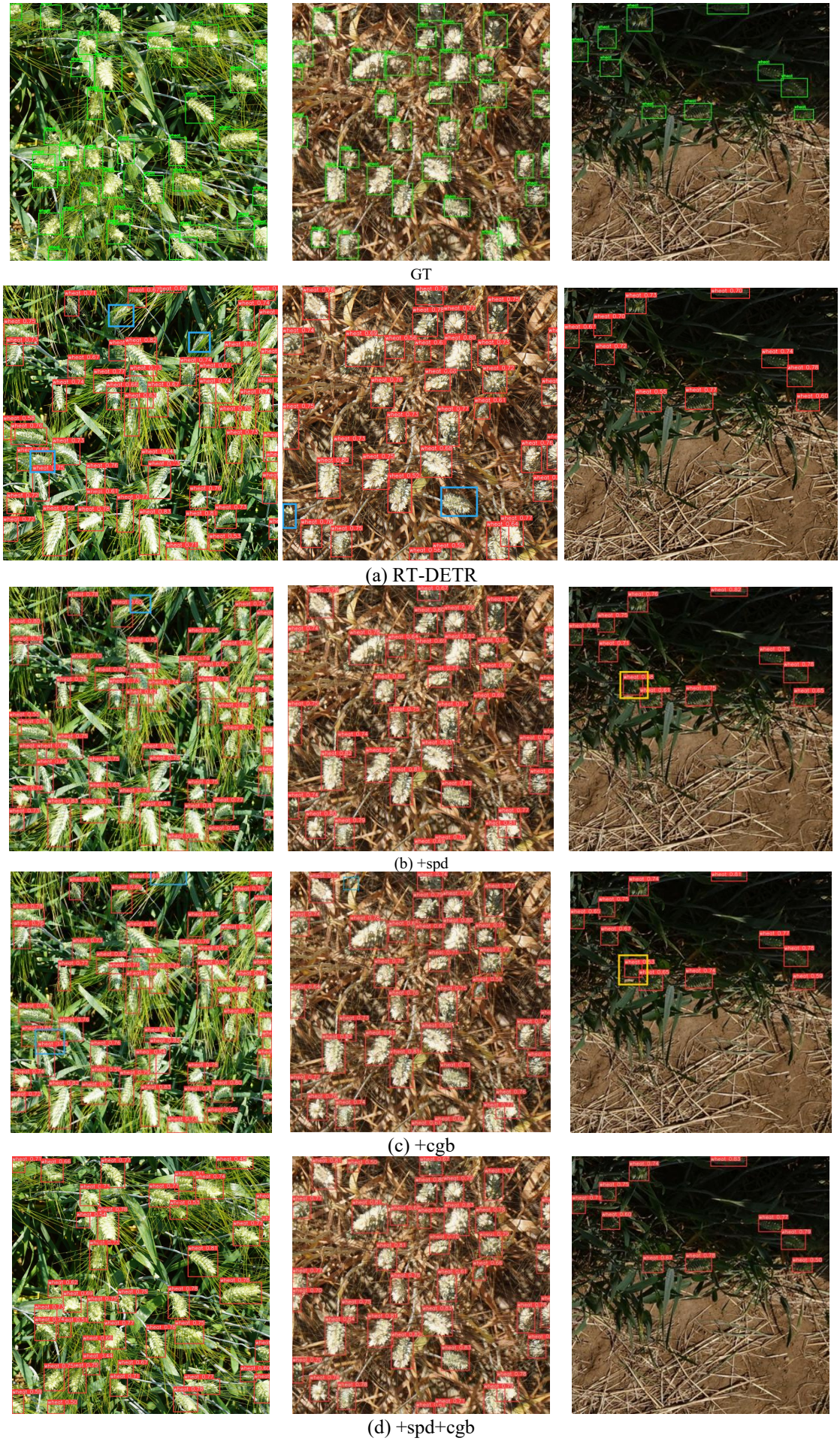


Fig.8. Comparison of visualization results of ablation experiments

TABLE III
DETECTION RESULTS OF FIELD WHEAT EARS BY CLASSIC AND ADVANCED DETECTION MODELS

Model	AP_{50}	AP_{75}	AP_{50-95}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L	Params/M
Faster R-CNN	89.6	47.9	48.8	11.2	47.9	55.7	13.1	54.0	61.2	41.34
Retinanet	85.5	48.0	47.9	3.7	46.9	56.8	3.7	52.8	62.3	36.33
FoveaBox	91.5	47.3	49.4	10.7	48.8	55.5	13.2	56.1	61.4	36.24
YOLOX	90.6	45.8	47.8	14.1	47.5	54.2	26.6	55.7	60.4	8.93
Swin-Transformer	85.5	47.6	47.7	3.0	46.7	56.9	2.8	52.7	62.6	36.82
Pyramid Vision Transformer	85.3	46.2	46.9	2.8	46.2	55.2	2.7	52.4	61.5	21.33
RT-DETR	91.9	50.1	50.7	16.2	48.6	55.0	1.6	14.5	55.7	32.78
Ours(Improved RT-DETR)	93.5	56.2	54.4	22.2	52.0	59.4	29.7	58.9	65.3	38.94

the blue-highlighted box signifies the overlooked target.

As illustrated in Figure 8, the baseline model exhibits insensitivity to identifying wheat ears in overlapping areas, cluttered backgrounds, and along the edges of images, resulting in instances of missed detections. The implementation of the SPD-Conv module mitigates the problem of missing detections in the model. However, due to an excessively strong learning capability, non-wheat ear regions are misidentified as wheat ears, leading to occurrences of false detections. When the CG Block module is introduced in isolation, the model still encounters missed detection issues under overlapping and complex background conditions, although there are significant improvements compared to the baseline model. Furthermore, when both modules are applied to the model, the optimized version demonstrates enhanced accuracy in detecting wheat ears in overlapping areas, complex backgrounds, and along the edges of images, without any instances of false detections. This presents compelling proof that the proposed model in this research can proficiently execute wheat ear detection, thus enabling precise assessments of wheat output.

E. Comparative experiments

This research chooses the most sophisticated object detection algorithms for comparison in order to better examine effectiveness of the model in identifying wheat ears. The comparative experiments and those conducted in this study are executed in an identical environment, with parameter settings aligned with the model presented herein to guarantee the integrity of the comparative analysis. The comparative models include two-stage object detector Faster R-CNN, single-stage object detector Retinanet, single-stage anchor free box algorithm FoveaBox [32], YOLOX for lightweight backbone networks in the YOLO series [33], image classification models Swin Transformer [34] and Pyramid Vision Transformer [35] under the Transformer architecture. Table 3 illustrates the detection efficacy of each model employed in the comparative experiment.

The experimental findings shown in Table 3 distinctly illustrate the performance of each model on the GWHD2021 dataset. The optimized model derived from RT-DETR surpasses other leading methodologies in this investigation. Specifically, the average accuracy of the model is 93.5% at an IoU threshold of 0.5, and 56.2% as the IoU threshold is increased to 0.75. The GWHD2021 dataset comprises 2,799 training sets and 700 validation sets, featuring 1,568 small-scale targets in the training set, which constitutes 1.2% of the total, and 491 small-scale targets in the validation set, representing 1.52% of that set. The proportions of these scales are detailed in Table 1. Due to the insufficient amount of small-scale wheat ear data, the model is inadequately trained for small-scale targets, resulting in lower detection accuracy. The AP values for small-scale targets in the comparative tests are often low, whereas the improved RT-DETR model consistently exhibits high detection accuracy, suggesting that the optimized model possesses superior generalization capability and robustness in small-scale target detection tasks. The optimized model achieves Average Recall Scores (AR_S), Average Recall Metrics (AR_M), and Average Recall Levels (AR_L) of 29.7%, 58.9%, and 65.3%, respectively, all of which surpass the performance of the other comparative models.

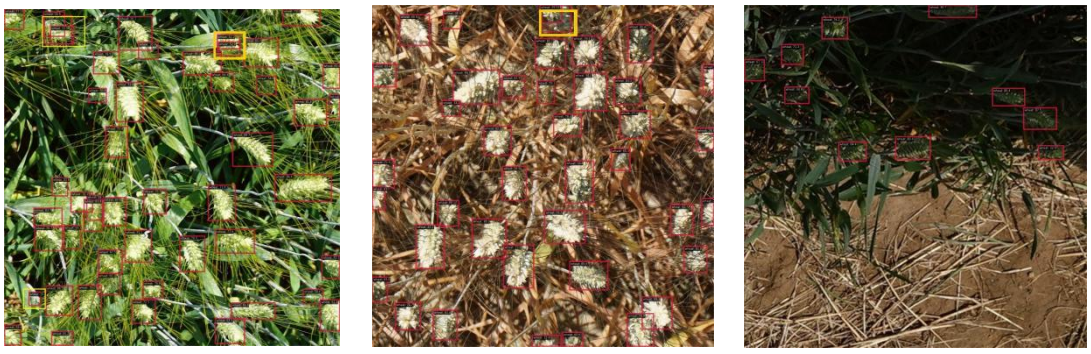
Considering the parameter count, the optimized model has a parameter count of 38.94M. Because of the characteristics of RT-DETR model its Transformer architecture itself, such as the self-attention mechanism and the multi-head attention mechanism which require a large number of weight matrices, large-scale input embedding and positional encoding, these designs result in the RT-DETR model possessing a substantial number of parameters. Notwithstanding the numerous covariates in the RT-DETR model, the optimized version displays formidable detection proficiency regarding accuracy and recall, particularly showcasing substantial interference resistance and robustness while addressing objects of varying scales. Overall, the optimized RT-DETR model demonstrates superior results across each evaluation metric on the GWHD2021 dataset. Figure 7 demonstrates the comparison of the detected images.

The optimized RT-DETR model outperforms seven other classic models on the Global Wheat Spike Detection dataset. As illustrated in Figure 8, the optimized model can accurately detect wheat ears without any instances of missed

GT

Faster R-CNN

Retinanet



FoveaBox

YOLOX

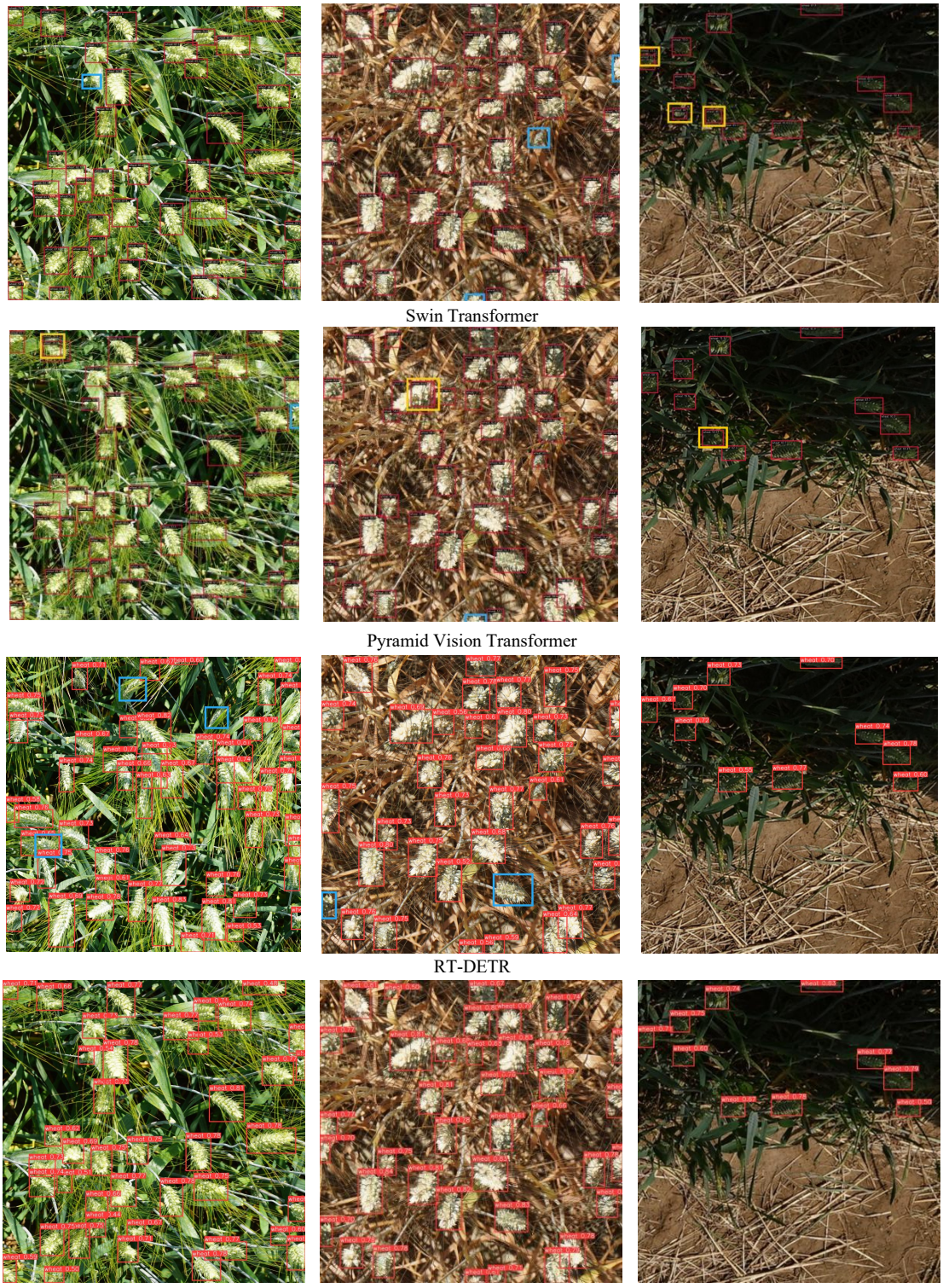


Fig 9 Comparison of detection images of global wheat ear data set (yellow box marks the misdetection area; blue box marks the missed detection target)

detections or false positives. In contrast, Faster R-CNN is prone to false positives in scenarios with a high degree of overlap among wheat ears; RetinaNet encounters missing detections and occasional false positives in peripheral areas; FoveaBox exhibits false detections in overlapping and complex scenes; YOLOX also encounters false positives under dense conditions; Swin Transformer is susceptible to missed detections in overlapping and complex backgrounds, as well as false positives due to shadow occlusion; similarly, the Pyramid Vision Transformer model shows missed detections under dense and shadow conditions. Enlarge specific areas of the wheat ear images to more clearly illustrate the detection performance differences between the comparison images and the model results, as depicted in Figure 10. The red box highlights the complex regions in the original image, the blue box signifies the detection outcomes of the comparison model, while the yellow box indicates the outcomes of detection of the improved RT-DETR model. It is evident that in crowded, obstructed, and darkened situations, the comparison model shows missed detections

Fig.10. Detail image of wheat ear detection in comparative experiment

with fewer anchor boxes and erroneous detections with more anchor boxes. In contrast, the model developed in this study attains precise detection grounded on the actual conditions of the wheat ears. Overall, the optimized RT-DETR model demonstrates exceptional detection performance across various complex environment.

IV. Conclusion

This study optimizes the RT-DETR model using the GWHD2021 dataset to address the target identification issue caused by the complicated growth environment and dense growth state among wheat ears, which significantly increases the accuracy of target recognition of wheat ears. On the one hand, this paper proposes to introduce the SPD-Conv module in the backbone network part to strengthen the ability of capturing spatial details of the features and make up for the loss of information in the backbone due to DWConv, so as to better detect the overlapping part. On the other hand, the Conv operation in the hybrid encoder part is replaced with Context Guided Block to strengthen the connection between contexts, hence improving the exactness of wheat ear detection in intricate conditions and augmenting the generalization ability of the model.

The optimized RT-DETR model presented in this study effectively achieves accurate detection of wheat targets with a model accuracy of 91.2% and an AP_{50} of 93.5%. Relative to other sophisticated target detection models, the approach presented in this paper sustains elevated accuracy. which can help farmers to know the yield of wheat field in time and provide accurate data support for agricultural production, thus promoting the development of agriculture. Although the higher parameter count of the model restricts its deployment and application in resource-constrained environments, the improved feature extraction and implementation of advanced detection algorithms significantly boost the effectiveness of small target detection, overcoming the shortcomings of traditional methods in identifying small targets. Future research must achieve the lightweighting of the target detection model while maintaining precision in detection, to acknowledge the extensive utilization of target detection technology in agriculture and to advance the progression of precision agriculture.

References

[1] Curtis T, Halford NG. Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. Annals of Applied Biology, 2014, 164(3):354-372.

[2] Zhou X, Zheng H, Xu X, et al. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, 130: 246-255.

[3] Xu X, Li H, Yin F, et al. Wheat ear counting using K-means clustering segmentation and convolutional neural network. Plant Methods, 2020, 16: 1-13.

[4] Tian H, Wang T, Liu Y, et al. Computer vision technology in agricultural automation —a review. Information Processing in Agriculture, 2019, 7(1): 1-19.

[5] Sadeghi-Tehran P, Virlet N, Ampe E, et al. DeepCount: in-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks. Frontiers in Plant Science, 2019, 10:1176.

[6] He H, Li Z, Tian G, et al. Towards accurate dense pedestrian detection via occlusion-prediction aware label assignment and hierarchical-NMS. Pattern Recognition Letters, 2023, 174: 78-84.

[7] Fu B, Li W, Sun Y, et al. Correlated NMS: establishing correlations between dense predictions of remote sensing images. IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2023: 6153-6156.

[8] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7774-7783.

[9] Du J, Liu L, Li R, et al. Towards densely clustered tiny pest detection in the wild environment. Neurocomputing, 2022, 490: 400-412.

[10] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning. Journal of Big Data, 2019, 6(1): 1-48.

[11] Liu X, Li G, Chen W, et al. Detection of dense Citrus fruits by combining coordinated attention and cross-scale connection with weighted feature fusion. Applied Sciences, 2022, 12(13): 6600.
[12] Hou Q, Zhou D, Feng J. Coordinate Attention for Efficient Mobile Network Design. 2021, 10.48550/arXiv.2103.02907.
[13] Tan M, Pang R, Le QV. EfficientDet: Scalable and efficient object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020.
[14] Wen X, Yao Y, Cai Y, et al. A Lightweight ST-YOLO based model for detection of tea bud in unstructured natural environments. IAENG International Journal of Applied Mathematics, vol. 54, no. 3, pp 342-349, 2024.
[15] Li L, Hassan M A, Yang S, et al. Development of image-based wheat spike counter through a Faster R-CNN algorithm and application for genetic studies. The Crop Journal, 2022, 10(5): 1303-1311.
[16] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28: 91-99.
[17] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023, 111(3): 257-276.
[18] Wen C, Wu J, Chen H, et al. Wheat spike detection and counting in the field based on SpikeRetinaNet. Frontiers in Plant Science, 2022, 13: 821717.
[19] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[20] Law H, Deng J. Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.
[21] Wang M, Sun K, Guo A. Wheat ear detection using anchor-free ObjectBox model with attention mechanism. Signal, Image and Video Processing, 2023, 17(7): 3425-3432.
[22] Zand M, Etemad A, Greenspan M. Objectbox: From centers to boxes for anchor-free object detection. European Conference on Computer Vision, 2022: 390-406.
[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
[24] Zhou Q, Huang Z, Zheng S, et al. A wheat spike detection method based on Transformer. Frontiers in Plant Science, 2022, 13: 1023924.
[25] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[J]. arXiv preprint arXiv:2304.08069, 2023.
[26] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. European conference on computer vision, 2020: 213-229.
[27] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2022: 443-459.
[28] Wu T, Tang S, Zhang R, et al. Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.
[29] Zhang P, Lo E, Lu B. High performance depthwise and pointwise convolutions on mobile devices. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6795-6802.
[30] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 658-666.
[31] David E, Serouart M, Smith D, et al. Global wheat head dataset 2021: more diversity to improve the benchmarking of wheat head localization methods. arXiv preprint arXiv:2105.07660, 2021.
[32] Kong T, Sun F, Liu H, et al. Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.
[33] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.
[34] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
[35] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578