IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 January 2023, accepted 30 January 2023, date of publication 1 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241808

RESEARCH ARTICLE

YOLOv5-Based Model Integrating Separable Convolutions for Detection of Wheat Head Images

RAN SHEN, TONG ZHEN, AND ZHIHUI LI

College of Information Science and Engineering, Henan University of Technology, Zhengzhou, Henan 450001, China
Key Laboratory of Grain Information Processing and Control, Ministry of Education, Henan University of Technology, Zhengzhou, Henan 450001, China

Corresponding author: Tong Zhen (zt@haut.edu.cn)

ABSTRACT In the detection of global wheat heads, it is easy to give rise to difficulties due to different wheat varieties, planting densities and growth periods of wheat plants in different countries. In addition, the illumination conditions of the image collection and the complex background of field will also reduce the detection accuracy. It is also hard to accurately detect targets that are occluded and partially displayed in the image. To solve the above problems, in this paper, an improved YOLOv5 algorithm that integrates separable convolution and attention mechanisms is proposed. Firstly, the number of CSP modules of YOLOv5 is reduced to shrink memory consumption. Subsequently, vanilla convolutions in the CSP are replaced by separable convolutions which is also added to the fusion path and to reduce the redundant information of the feature map, so as to reduce the complexity of the model. In addition, the co-attention mechanism is added in backbone. Finally, the feature fusion module was adjusted to make the high-level features fuse more low-level information. Compared with the original algorithm, results show that the mAP of the improved algorithm reaches 93.8% which is 4.2% higher than that of the YOLOv5 algorithm, and the FPS is 27.4 which is 1.3 higher than YOLOv5. YOLOv7 is emphatically compared during model evaluation, other YOLO series and mainstream detection algorithms are also compared, and results show that our model has the best inference time and the best accuracy when dealing with high pixel images.

INDEX TERMS YOLOv5, separable convolution, feature fusion, object detection, attention mechanism.

I. INTRODUCTION

Wheat production is related to the country's food security. The prediction of wheat yield forecast can provide references for agricultural production management decision-making, and also provide support for the government's macroeconomic regulation of rural land policies and food prices. When farmers make management decisions in the field, the health and maturity of wheat can be assessed by estimating the density and size of different varieties of wheat heads based on images of ''wheat heads'' -the spikes at the top of plants containing grains. In fact, in the field, the yield of wheat is mainly related to the number of wheat ears per unit area, the thousand seed weight and the number of grains per ear. Among them, the number of wheat ears per unit area is the most important index, which directly reflects the growth status and quality of wheat [1]. However, accurate detection of wheat heads in outdoor field images can be challenging in computer vision. Dense wheat plants often overlap, and wind can blur images. Both make it difficult to identify individual heads. In addition, appearance varies by maturity, color, genotype, and head orientation. Finally, since wheat is distributed all over the world, different varieties, planting densities, patterns, and field conditions must be taken into account. Models developed for wheat phenotype need to generalize across different growing environments.

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

TABLE 1. Research and analysis of the latest relevant literature.

Paper	Model	Advantages	Limitations
Feixiang et al	YOLOv5+WCA	The weighted average makes the extracted features richer.	The WCA module will decrease the speed of detection.
Alexis et al[10]	SLIC	It provides a reference for the application of segmentation method in wheat head detection.	Results of post-flowering is bad.
Yan et al[11]	SPSA + Shuffle Units	It is lightweight and has good reasoning effect.	The model generalizes weakly and is prone to overfitting.
Chengxin et al	DCT	It has particularly good robustness to targets under different illumination conditions.	It is difficult to distinguish objects that are similar to the background color.
Sandesh et al[12]	MDWConv + MSPP	The hybrid deep convolution simplifies multiple convolution kernels into a single convolution kernel, which can effectively improve the inference speed.	Although the total amount of computation is smaller, the number of layers is larger, so the improvement of time is not obvious for highly parallel devices.

At present, there are mainly three mainstream methods for the identification and counting of wheat heads: image processing, machine learning and deep learning.

In the study based on image processing, Fernandez-Gallego [2] used Laplacian filter and median filter to remove noise, and then used local maximum peak for wheat head counting. A year later, he experimented with thermal imaging, using contrast enhancement and filtering to count wheat heads with 90 percent accuracy. Zhou et al. [3]. used multiple sensors to integrate the improved maximum entropy segmentation algorithm to identify wheat heads. He first used Gram-Schmidt fusion algorithm to fuse multispectral images and panchromatic images, then used maximum entropy segmentation method for coarse segmentation, and finally used morphological reconstruction theory to segment the adhesion part for fine segmentation. However, the problem of image edge distortion will lead to the reduction of recognition accuracy. At the same time, the resolution of the collected image has a great impact on the segmentation effect of sticky wheat.

In the research of wheat heads recognition based on machine learning, Alharbi et al. [4]. used Gabor filter bank and K-means clustering algorithm to detect wheat heads, and the average accuracy reached 90.7%, but the overall performance was not too high, because the calculation time is too long and storage space consumption is too large when extracting texture features. In order to make the model more robust to images of wheat head in different growth stages and different weather, Sadeghi-Tehran [5] proposed an efficient computing system called DeepCount, which contributed to the high-throughput analysis of wheat ears in the field. However, this model has high requirements on the collected images, for example, it needs to collect images that are on the level ground and have the same sample area.

In addition to the above two methods, more scholars use deep learning for recognition of wheat heads. Samadur [6] reproduced the YOLOv5 algorithm using the same data set as ours, which laid a foundation for the later application of the YOLOv5 algorithm on the wheat data set. However, the paper did not further modify the YOLOv5 model, so the accuracy and speed were not greatly improved, and it was hardly innovative. Saeed [7] uses MobileNet as a lightweight backbone, and then uses the feature maps output from the backbone as the input of counting network branches and positioning network branches respectively. Compared with other networks, Saeed has fewer parameters, which is easy to deploy on mobile terminals and conducive to farmers' field use. Feixiang [8] used YOLOv5, weighted coordinate attention mechanism and image processing technology to achieve wheat head detection, and used weighted average to replace the simple arithmetic average in CA attention mechanism. The purpose is to solve the universality of the previous algorithm and improve the generalization ability, but the author did not describe the reasoning speed of the modified model, and the inference accuracy of the model is only 0.862. Chengxin [9] uses a dynamic color transformation network to detect wheat heads, which can adapt to illumination changes in the image and does not require additional memory overhead, and won the runner-up in the 2021 Global Wheat Challenge. However, the author also mentioned that the network cannot detect wheat heads with similar color to the background. Moreover, TABLE 1 also sum up the newest and typical models with analyzing the benefits and limitations.

It can be seen from the above methods that each research methods have their own shortcomings. For example, when image processing technology is used for detection alone, the traditional segmentation methods, such as watershed [13], concave point segmentation [14] and corner detection [15], have poor processing effect on a large number of adhesion and occlusion wheat heads, so the overall recognition accuracy is reduced. The combination of image processing and machine learning can overcome the interference of different backgrounds and different lighting environments on the experimental results, but it is easy to be affected by image noise. Deep learning methods can effectively solve the above problems, and there are many studies using deep learning for identification and counting of wheat heads, However, this good effect is very limited. In other words, first of all, the

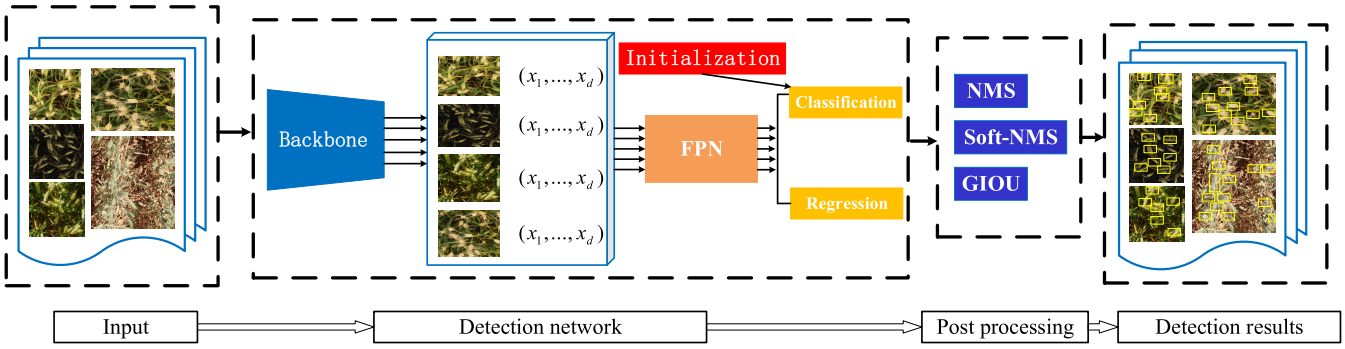

FIGURE 1. Flow chart of wheat head test.

data set used by these models is single and there are few cases of multiple wheat ears adhesion. Moreover, they all need to take the direction when collecting images and keep the ground flat where the wheat ears are which leads to a poor recognition effect for wheat heads under the concave and convex ground in the actual situation. In addition, these models do not consider the presence of a part of the wheat heads in the collected image.

In addition, all the above research methods have four common and fatal limitations, as follows:

1. Complex background and different illumination will greatly affect the recognition accuracy.
2. No good treatment of overlapping, adhesive and occluded wheat heads.
3. Most of the model validation results are based on the data set of specific regions, and the universality is not high.
4. Fast recognition depends on hardware. For general hardware, the performance is not greatly improved, that is to say, a large part of the improved performance of the improved model is due to the advantages of the hardware itself.

In view of the above problems, this paper proposes a method for identifying and counting wheat heads in the field based on YOLOv5. The structure of this paper is as follows: The second part introduces the basic framework and improved method of the model we use. The third section describes the experimental process, and analyzes, evaluates and discusses the experimental results. The fourth part makes a phased summary of the existing work, summarizes the limitations of our current research and puts forward the prospects.

II. MATERIALS AND METHODS

A. WORKFLOW OF WHEAT HEAD DETECTION

We simplify the research work as a flowchart shown in Fig.1. In the actual research process, we perform mosaic data augmentation on the input image, and then input the enhanced image into the feature extraction network to extract the features of the target, which are actually many multi-dimensional tensors. Before the operations of classification and regression, a feature pyramid network is added to fuse the low-level information including location, color, texture et al. with the high-level information to enhance the positioning ability of the model. At the end of the detection network, the fused features were inputted into the branch of classification and regression respectively for target category recognition and locating. After successful target recognition and localization, post-processing such as non-maximum suppression is performed to screen the target boxes with low scores. Finally, the target box is displayed on the original image as output results.

B. MODEL STRUCTURE OF YOLOv5

Before a series of YOLO models were proposed, the algorithms based on the RCNN series dominated the field of object detection. Although the RCNN series models had high detection accuracy, they were criticized for their detection rate that could not meet real-time performance. The idea of YOLOv1 [16] is to divide an image into a number of grids, and if the center of an object falls in this grid, this grid is responsible for predicting that object. Compared with the two-stage Faster RCNN model, YOLOv1 has good inference speed but poor accuracy. Compared with the SSD model, YOLOv1 has no advantage, and the predicted effect of densely distributed small targets is poor. However, the YOLOv1 model still has something to learn from. In 2017, Joseph proposed YOLOv2 [17], which added a BN layer on the basis of YOLOv1, adopted a higher resolution classifier, and used the Anchor-based target bounding box prediction method to improve performance and accuracy. In 2018, YOLOv3 [18] was proposed, which uses Darknet53 as the backbone. The improved YOLOv3_SPP network based on YOLOv3 performs well. It uses Mosaic data augmentation and adds Spatial Pyramid Pooling (SPP) module. The feature fusion of different scales is realized. In addition, CIOU is used to calculate the loss, so that the model can converge faster, and the positioning accuracy is higher. In 2020, YOLOv4 [19] modified the backbone on the basis of YOLOv3_spp, added Path Aggregation Network (PAN), and also re-optimized anchor, which further improved the inference speed and detection accuracy. In the same year, YOLOv5 optimized the

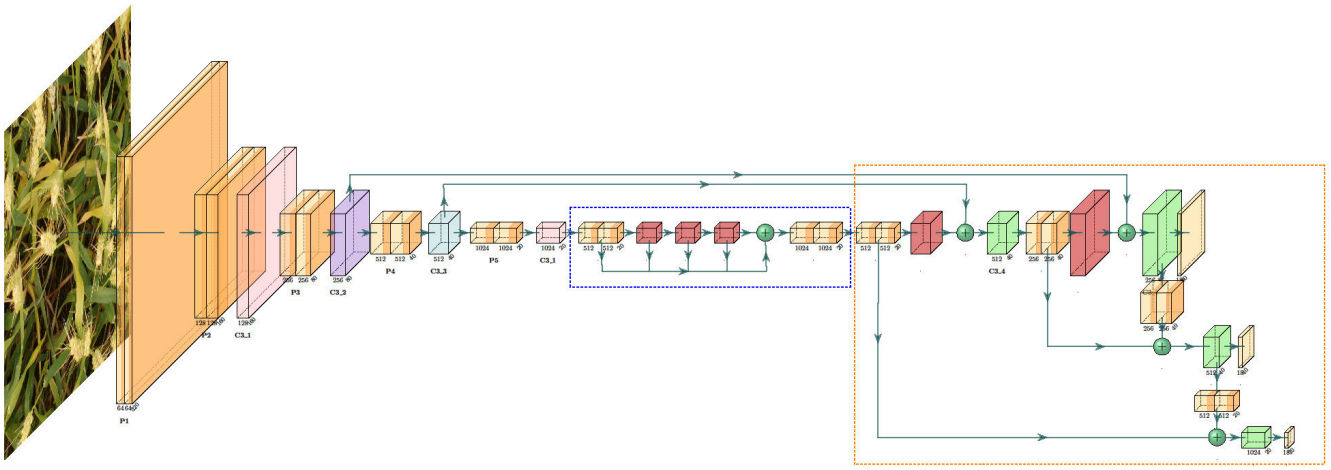

FIGURE 2. The Structure of YOLOv5.

backbone and SPP module on the basis of YOLOv4, which further improved the performance and accuracy.

Compared with YOLOv4, YOLOv5 has the characteristics of smaller mean weight file, shorter training time and faster inference speed on the basis of a small reduction in detection accuracy. The YOLOv5 network structure mainly includes four parts: Input, Backbone, Neck and Head. The input is mainly Mosaic data augmentation, adaptive anchor box calculation, and adaptive image scaling. Mosaic data enhancement stitches four images by random scaling, random cropping, and random arrangement, which enriches the background and small target of the detected targets and makes the network more robust. The principle of adaptive anchor box calculation is to output the predicted box on the basis of the initial anchor box, and then compare it with the real box, calculate the gap between the two, and then update the network parameters in the opposite direction, and iterate the network parameters to finally obtain the best anchor box information. Adaptive image scaling refers to the adaptive addition of black edges to the original image, uniformly scaling to a standard size. Backbone mainly includes Focus and CSP structures. Focus structure slices the image, quadrupling the channel of the image, half of the width and length of the image, and obtains the feature map after a convolution. The CSP structure is to divide the data into two parts, and one part is convolved, while the other part is not processed, and then the results of the processed two parts are concatenated. The use of CSP module reduces the memory consumption and enhances the learning ability of CNN. The feature pyramid networks (FPN [20]) and path aggregation network (PAN [21]) modules are integrated in Neck. The FPN layer conveys strong semantic features from top to bottom, and the feature pyramid from bottom to top conveys strong localization features. The backbone layer and the detection layer were integrated to make the model obtain richer feature information. Head outputs a tensor with object class probabilities, scores, and bounding box information. It consists of three detection layers, and features of different scales are used to detect objects of different sizes. The model structure of YOLOv5l is shown in Fig.2, where the green cube represents the modules fused with the CSP structure. C3_1 contains three BottleNeck1 modules, C3_2 contains six BottleNeck1 modules, and C3_3 contains nine BottleNeck1 modules. C3_4 contains three BottleNeck2 modules. The structure of the BottleNeck1 and BottleNeck2 modules are shown in the dashed box in Fig.4. The module inside the blue dashed box in Fig.2 is the SPPF module, the orange dashed box is the PAN module, and the plus sign in the figure refers to concatenation.

YOLOv7 is the latest version of yolo series. Compared with the previous YOLO, it mainly improves some of the structures, but there is no great update in the whole framework. It can still be divided into three parts, namely backbone, neck and head. We focus on comparing the YOLOv5 and YOLOv7 models, as shown in Fig.3. Compared with YOLOv5, the backbone of YOLOv7 adds multi-branch stacking module (E-ELAN) and transition module (MPConv) to replace C3 module and ordinary convolution in YOLOv5, as well as replaces SPPF module in YOLOv5 with SPPCSPC module. The SPPCSPC module has one more branch of convolution operation and one more concatenation operation than SPPF. In neck, PANet is used as in YOLOv5. The RepConv structure originated from RepVGG [22] is added before the head module, which combines the convolution operation and the normalization operation into one convolution operation to achieve the effect of improving the inference speed of the model.

C. PROBLEMS OF THE ORIGINAL MODEL IN WHEAT HEAD DETECTION

Although YOLOv5 is far superior to other models in terms of inference speed and accuracy, there are still the following defects for the wheat head image detection problems that need to be solved in this paper.

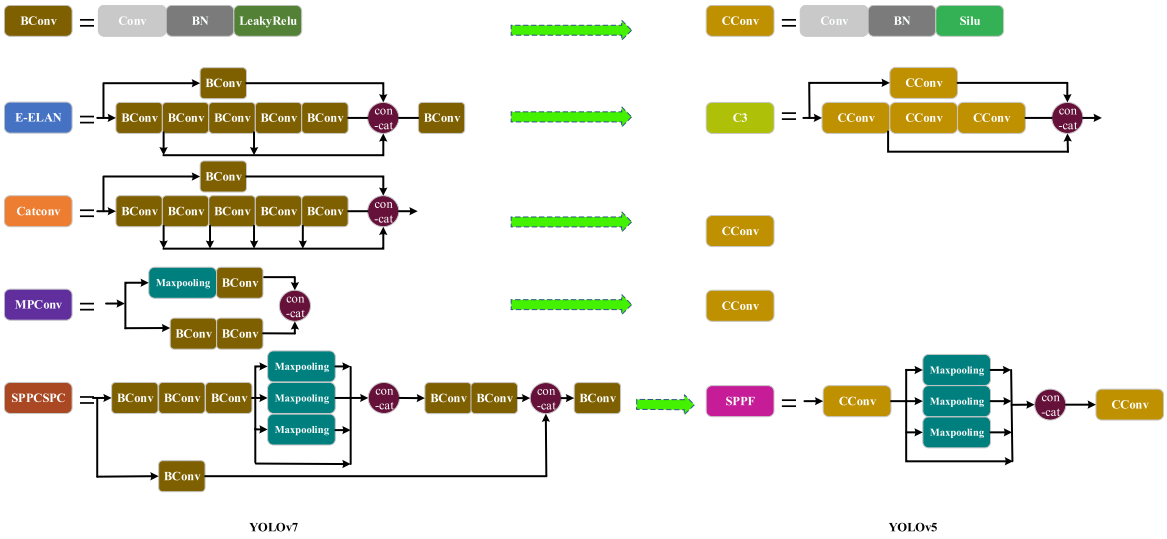

FIGURE 3. The contrast between YOLOv5 and YOLOv7(module to module).

1. The Bottleneck structure contains many convolution operations, and the kernel contains a lot of parameters, which leads to a lot of memory consumption when deploying the model. More features of the shallow layers are needed to deal with wheat ear adhesion and occlusion, and it is easy to cause the shallow information to be blurred or completely lost after a large number of convolutions, which undoubtedly reduces the localizability of the model.
2. In the original model, the feature maps sampled by 8 times and 16 times of 32 are used as the feature layer to detect the target. When the input image size is 640 $\times$ 640, the detection layers of 80 $\times$ 80, 40 $\times$ 40 and 20 $\times$ 20 can be used to detect the target size above 8 $\times$ 8, 16 $\times$ 16 and 32 $\times$ 32 respectively. Therefore, it will be difficult to detect an object when its size is smaller than 8 $\times$ 8 in the original input image.
3. The complex background is easy to reduce the feature extraction ability of the model, so a strategy is needed to improve the attention span of the model in the detection process, so that the model can accurately extract the target features that need to be extracted.
4. Each feature layer extracted by the network contains many feature maps, and there are similar patterns between these feature maps, that is, redundant information, which will cause a large amount of calculation of the network, and then reduce the inference speed. In addition, because the feature maps of the same layer are similar but not identical, it is difficult to determine which feature map to eliminate and whether the feature map to eliminate contains important detail information, so it is impossible to directly eliminate redundant features.

D. IMPROVEMENT OF MODEL

1) OPTIMIZATION OF THE CSP

In the backbone of the original model, the Bottleneck1 module repeats 3, 6, 9, and 3 times, respectively, and processes the features separately. One branch for convolution, while the other branch is fed into the CSP module, and then the two branches are integrated. However, the disadvantage of this method is that with the continuous increase of convolution operations, on the one hand, the amount of calculation increases dramatically, on the other hand, the underlying information such as edge features, texture features, position information of the target becomes less, which is not conducive to the recognition of mutually occluded wheat ears. Therefore, in order to reduce the amount of computation and extract more shallow information, this paper reduces the number of iterations of the Bottleneck1 module from 3, 6, 9, and 3 to 3, 3, 3, respectively. At the same time, it removes the convolution operation of the branch that does not add the Bottleneck1 module. We directly let the feature merge with the feature from the branch output that has the Bottleneck1 module.

2) INTRODUCTION OF COLLABORATIVE ATTENTION MECHANISM

The attention mechanism can make the model obtain more location information, improve the feature extraction ability of the model, and help the model to identify the target of interest more effectively. Therefore, adding the attention mechanism can improve the recognition ability of the model for wheat heads in complex backgrounds. Attention mechanism includes channel attention mechanism, spatial attention mechanism, and the fusion of them. Commonly used attention mechanisms include ECA [23], CA [24], CBAM [25], SE [26]. The core idea of SE is to automatically learn feature weights based on loss through a fully connected network.

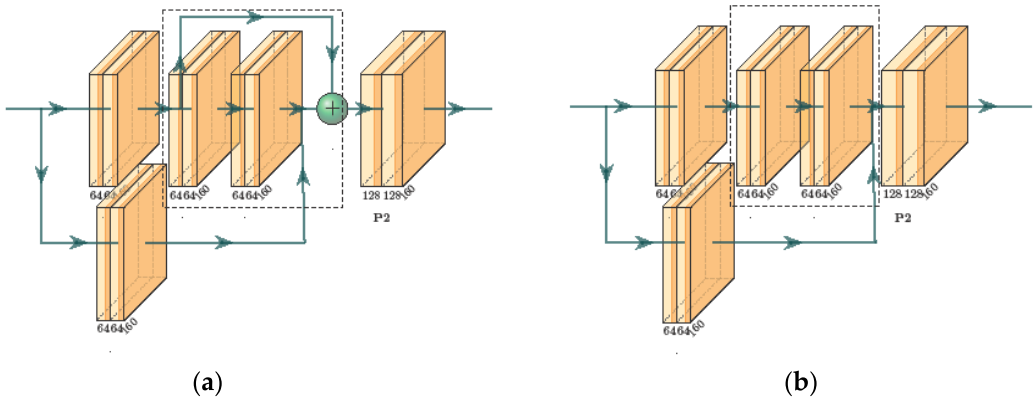

FIGURE 4. The structure of Bottleneck (a) The structure of the Bottleneck1; (b) The structure of the Bottleneck2.

Based on the SE module, ECA changed the use of a fully connected layer to a 1×1 convolutional layer in SE to learn channel information. CBAM combines channel attention and spatial attention modules, which can achieve better results compared with the attention mechanism that only focuses on channels. CA performs global average pooling on the two directions of height and width respectively, and then concatenates the feature maps of these two directions. CA can efficiently handle the relationship between channels. However, these methods increase the parameters of the model. To solve this problem, Yun et al. proposed a normalized attention mechanism, NAM [27], to highlight salient features by using the variance measurement of the training weights. It applies a coefficient weight penalty on the attention module, which can reduce the weights of less salient features, making these weights more computationally efficient while being able to maintain the same performance. In this paper, NAM module, ECA module, CA module, SE module and CBAM module are added respectively after backbone, and the models after adding various modules are compared in terms of reasoning speed and accuracy.

3) FEATURES FUSION BASED ON SE PARABLE CONVOLUTION

As shown in Fig.5, the image in the upper left corner is the original image of wheat head as input, and the other images are feature maps, among which there are many similar feature maps. Because the feature maps of each layer are similar but different, redundant features cannot be directly eliminated. SPConv [28] can effectively solve this problem and only needs a small amount of computation. It selects some representative feature maps to supplement the intrinsic information, and the remaining redundancy only needs to supplement with tiny and different details.

SPConv divides the input feature maps proportionally into two parts, and one is representative information and the other is redundant information. The representative part uses convolution operation with kernel $k \times k$ to provide intrinsic information, and the redundant part uses convolution operation with kernel 1×1 to provide tiny details. The matrix representation is shown in (1). Among them, the first half of the right side of the equation is the representative part, and the second half is the redundant part, in which y represents the output matrix, x represents the input feature matrix, L is the number of channels of the tensor inputted to the SPConv, and these L channels are split into two parts in a ratio of α,one for the representative applying group convolution and the other for the redundant applying point-by-point convolution.$W_{i,j}, j \in [1, \alpha L]$ represents the parameters of the group convolution with 3×3 kernel, $W_{i,j}, j \in [1, \alpha L + 1]$ represents the parameters of the convolution with 1×1 kernel, which performs point-by-point convolution on the remaining redundant features with the number of $(1 - \alpha)L$.

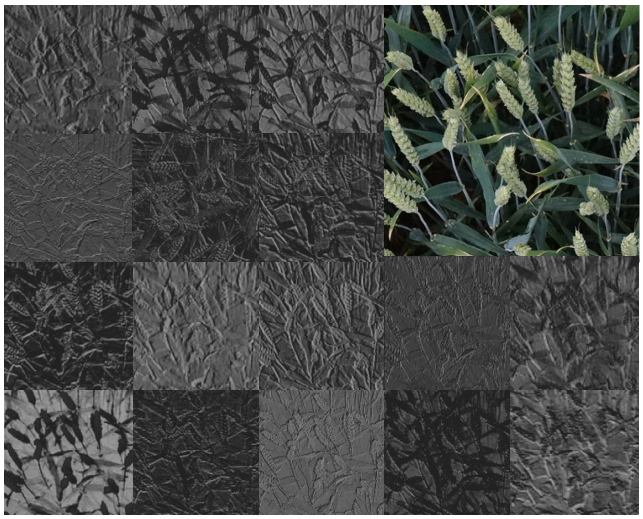

FIGURE 5. Visualization of partial feature maps.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1,\alpha L} \\ W_{21} & W_{22} & \cdots & W_{2,\alpha L} \\ \vdots & \vdots & & \vdots \\ W_{M,1} & W_{M,2} & \cdots & W_{M,\alpha L} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{\alpha L} \end{bmatrix}$$

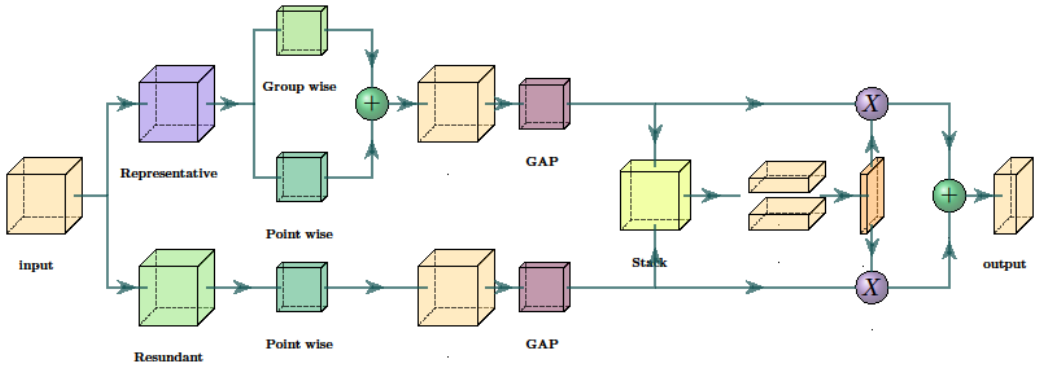

FIGURE 6. Schematic diagram of the SPConv model.

$$+\begin{bmatrix} W_{1,\alpha L+1} & W_{1,\alpha L+2} & \cdots & W_{1,L} \\ W_{2,\alpha L+1} & W_{2,\alpha L+2} & \cdots & W_{2,L} \\ \vdots & \vdots & & \vdots \\ W_{M,\alpha L+1} & W_{M,\alpha L+2} & \cdots & W_{M,L} \end{bmatrix} \begin{bmatrix} x_{\alpha L+1} \\ x_{\alpha L+2} \\ \vdots \\ x_L \end{bmatrix} \tag{1}$$

After dividing all the channels into two parts, the representative channels can be further divided, each part representing a class feature. On the representative part, there is still redundancy, and group convolution is used to reduce the redundancy. However, such an operation reduces the connection between channels, so a pointwise is added to compensate for this loss of information. Thus, the matrix representation of the representative part in (1) can be changed to (2).

$$\begin{bmatrix} W_{ll}^{p} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W_{GG}^{p} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_G \end{bmatrix} + \begin{bmatrix} \omega_{11} & \cdots & \omega_{1,\alpha L} \\ \vdots & \ddots & \vdots \\ \omega_{M,1} & \cdots & \omega_{M,\alpha L} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{\alpha L} \end{bmatrix} \tag{2}$$

In (2), W_{ii}^{p} is the grouped convolution parameter of the first group, $G$is the number of groups in which the representative channel is divided, and z_j is the number of channels included in the group j.

Then, the sum of channel-level statistics is generated by global average pooling (GAP) operation, and the two are stacked and compressed. Finally, the compressed feature vector is processed by the soft-max layer to regenerate two weighted feature vectors. The final output is obtained by multiplying and adding these two feature vectors with the representative part and the redundant part respectively, and the operation details are shown in Fig.6.

In the original model, the PAN module was concatenated with the output of the second and third CSP modules respectively, and the SPPF output features were also spliced with the PAN module. A separation-based convolution operation is added between the two CSP modules and the SPPF module and PAN module, so as to extract the representative information of the low-level features and make it fuse with the high-level features. In addition, the vanilla convolution in the CSP module is replaced by SPConv, so as to reduce the redundancy and calculation of the model and accelerate the inference speed of the model.

4) IMPROVEMENT OF PAN MODULE

In convolutional neural networks, the underlying features undergo fewer convolutions and retain the original contours, edges, colors, textures, angular and shape features. The position information is sufficient, and the target position is accurate, but because the receptive field is small, it is difficult to accurately recognize large targets, and the recognition accuracy of small targets is relatively high. High-level semantic features contain rich portfolio information and have strong ability to distinguish the target. However, due to the lack of low-level information such as location information, it is easy to cause inaccurate positioning. By analyzing the distribution of wheat ears in the field, it is found that two or more ears cover each other in most images, and there are also some ears that are not completely included in the image. Therefore, in order to further improve the positioning ability of the network, the low-level features and high-level semantic features can be efficiently fused to make it possess strong semantic information while still having strong perception ability for details.

In the original network, the output of layer 17 is beneficial to detect large objects, the output of layer 20 is beneficial to detect medium objects, and the feature map of layer 23 is beneficial to detect small objects. In order to enhance the perception of position information of the network, so as to enhance the detection of mutually occluded wheat ears, it is attempted to add a fusion layer to make up for the loss of spatial information caused by high-level feature resolution. There are four fusions in the original model, the first one is the fusion of the second CSP module (layer 4) and the first up sampling image after the SPPF module (layer 15), the second one is the fusion of the third CSP module (layer 6) and the second up sampling image after the SPPF module (layer 11), and the third and fourth are the fusion inside the PAN module. These are fusion at layer 10 and 21, and fusion at layer 14 and 18, respectively. In addition, this paper adds two more fusions: the fourth CSP module of the feature extraction layer

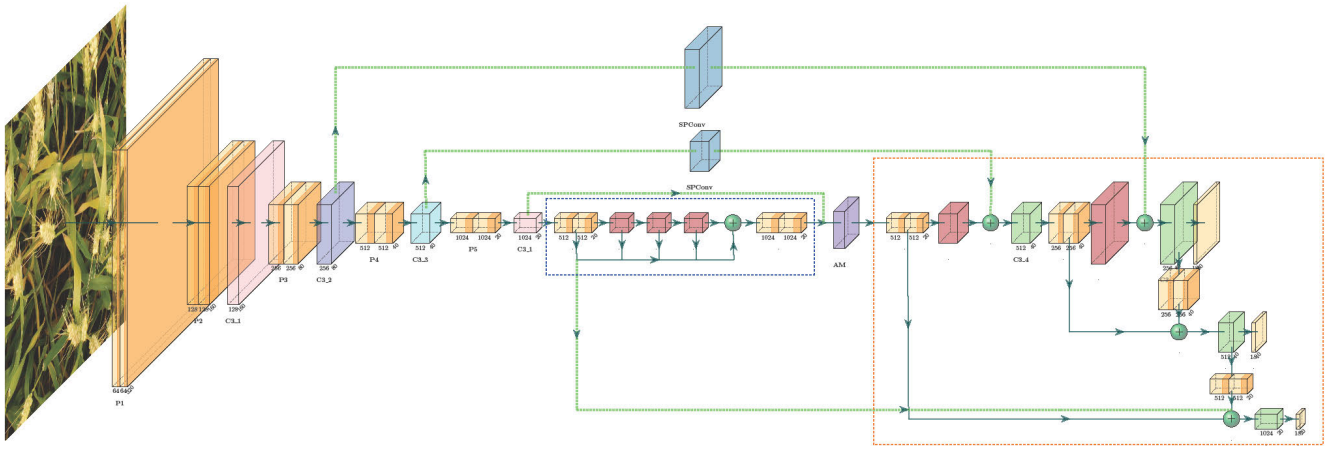

FIGURE 7. Structure of the improved YOLOv5.

FIGURE 8. Images of different wheat growth period. (a) post-flowering; (b) filling; (c) filling-ripening; (d) ripening.

(layer 8) will be integrated with the last convolution module of the SPPF module, and the first convolution layer of the SPPF module is concatenated with the 21st layer fusion. These two fusions fuse the output of the feature extraction layer with a large receptive field in the lower layer and the output of the feature extraction layer near the detection layer. The method of concatenation is used in all fusion operations to avoid information losses caused by the add operation.

The network structure diagram of the improved model is shown in Fig.7. The number of Bottleneck1 in each of the first three CSP modules has been reduced to three: and the green lines in the figure are the fusion routes we've added, the blue cubes are the SPConv modules, and the purple cubes are the attention modules. The improved yolov5 network structure in this paper can better adapt to the detection of wheat ears in complex scenes, and it has a certain improvement in the detection accuracy of wheat heads occluded and incomplete displayed.

III. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

The datasets in this paper are from the Global Wheat Head Detection (GWHD_2021 [29]) dataset. It comes from RGB images provided by a total of 19 institutions in 12 countries, with a total collection of 6500 images and 275,000 wheat heads. It has a high diversity, and includes images of various genotypes, different seeding densities and patterns, plant states and stages, and acquisition conditions. In total, the wheat heads images of post-flowering, filling, filling-ripening, and ripening are included, as shown in Fog.6. During the post-flowering period, the inner and outer glums of the flowers in the middle and upper part of the heads open, and the anther begins to powder. With green leaves and green spikes, the filling period is about 10 days after flowering, during which the wheat ears have green leaves and yellow spikes. During filling-ripening, the wheat leaves were yellow and green, and the spike was yellow. At the time of the harvest, the grain begins to harden, and farmers can harvest depending on the weather conditions. The leaves and spikes are yellow. In addition, images under different illumination conditions such as sunny day, cloudy day, backlight and beaming are also collected in this dataset, as shown in Fig.9.

Fig.10 visualizes the distribution of the size of the target box and the distribution of the location of the target box in the images. Fig.10(a) shows the coordinate distribution of the location center of the target box in the picture, where the coordinate is the relative position scaled according to the size of the picture. The horizontal axis is the x-coordinate of

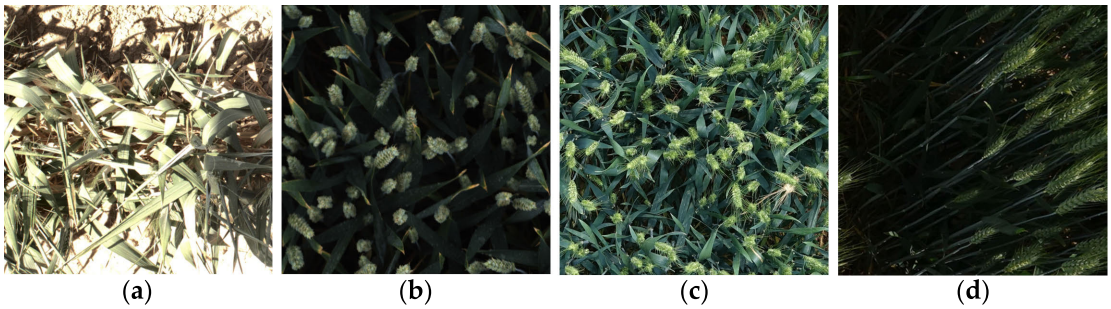

FIGURE 9. Wheat images under different lighting conditions. (a) Image captured under sunny and direct sunlight condition.; (b) Image captured under sunny and backlight condition; (c) Image captured under cloudy and direct sunlight condition; (d) Image captured under cloudy and backlight condition.

the center point of the target box after the image size is normalized to equal proportions, and the vertical axis represents the y-coordinate. Fig.10(b) shows the ratio of the length and width of the target box to the width and height of the whole image. The horizontal axis is the ratio of the target frame's width to the image's width, and the vertical axis is the ratio of the target frame's height to the image's height. The darkest color in the figure, the larger number of the targets are, and it can be seen from the two figures that the size of target boxes is not uniform, but the number of small targets is large, and the position of target boxes is evenly distributed in the whole picture.

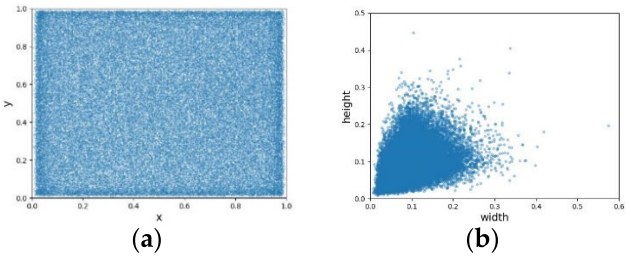

FIGURE 10. Distribution of bounding boxes in the dataset. (a) The location distribution of the bounding boxes; (b) Size distribution of the bounding boxes.

In addition, Mosaic data augmentation is applied to the images, and the Mosaic data augmentation method is an improvement of the CutMix [30] data augmentation method, which is based on stitching two images, and Mosaic is based on stitching four images. This operation not only increases the diversity of data and enhances the robustness of the model, but also makes the originally large target become a smaller target after being reduced, thereby reducing the over-response to large objects and enhancing the ability of the model to detect small targets, which solves the problem of small target detection in the data set to some extent. The results of data augmentation on one batch are shown in Fig.11.

FIGURE 11. Visualization of the results of Mosaic data augmentation.

2) MODEL EVALUATION METRICS

Common metrics for evaluating the performance of object detection models are FPS and GFLPOs, where FPS stands for the number of frames detected per second. The larger the value is, the better the performance of the model is. GFLPOs stands for one billion floating operations, and it measures the complexity and computation of the model. The lower the value is, the faster the model is. The evaluation indicators on accuracy include P, R, mAP@0.5, mAP@.5:0.95. P refers to the proportion of all samples predicted to be true, as shown in (3); R refers to the recall rate, which is the ratio of all samples predicted to be true, as shown in (4). Both values are expected to be high, and a larger P indicates a higher accuracy of detection; R indicates a more comprehensive detection. However, there are mutual restrictions between the two, so the P-R curve is often used to judge the accuracy of the model. AP (average precision) is the area of the P-R curve, the larger AP is, the higher accuracy of the model detection, mAP (mean average precision) is the mean of AP. It is common to use mAP@0.5 and mAP@0.5:0.95, where the former represents the area of the P-R curve at an IOU

of 0.5 and the latter represents the average area of all P-R curves at an IOU between 0.5 and 0.95. The main evaluation index in terms of memory consumption is parameters, and the larger the number of parameters, the more memory the model consumes.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

TP is the number of samples that predicted as a positive predicted as positive samples, and FN is the number of positive samples that are predicted as negative samples.

B. RESULTS

This paper designs three groups of comparative experiments, and all the experiments are implemented under Ubuntu18.04, 64-bit operating system, Tesla T4 graphics card, 15G video memory, PyTorch 1.10, CUDA10.1. The first group is to compare and analyze the influence of adding attention mechanisms in the model. The second group is to compare the performance and accuracy of the original YOLOv5 model and all the improvements. The last group is to compare the performance and accuracy of the improved model with all the YOLO series models and the current mainstream object detection algorithms.

The basic parameters are set as follows: epoch is set to 50, batch size is set to 16, the initial learning rate is 0.01, and warm-up and Mosaic data augmentation are both used for training. Loss variation during training is shown in Fig.12. It can be seen from the figure that the loss value decreases rapidly in the early stage of training, and with the increase of training epochs, the training loss gradually decreases and fluctuates around a boundary value. When epoch is 40, the loss decreases to a stable level, the model converges, and there is no overfitting during training.

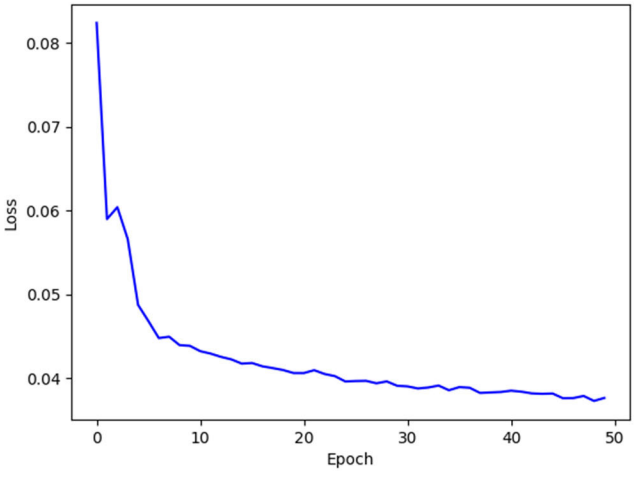

FIGURE 12. Loss in training.

1) IMPACTS OF ADDING DIFFERENT ATTENTION MECHANISMS ON THE MODEL

In this paper, we try to add five attention mechanisms, SE, ECA, CA, NAM, and CBAM, respectively after backbone, and design five groups of experiments to compare the performance and accuracy of the models after adding these five modules. TABLE 2 shows the results of the influence of adding different attention mechanisms in the model. As can sample, and FP is the number of negative samples that are be seen from the table, after adding the attention mechanism, the complexity of the model remains basically the same, the detection speed is slightly improved, and the accuracy can be improved by about 3%. Comparing the five attention mechanisms, it is suggested that CA has lower computational complexity and higher detection accuracy, yet after adding the CBAM module, the model accuracy decreases with the parameters increasing.

In addition, we use Grad-CAM to draw the heat maps of the original model and the model after adding the attention module, so as to visualize the key areas that the network pays attention to in the identification of wheat heads. The results are shown in Fig.13. In the figure, rows represent the visualization of detection layers, and column represents the results of the three detection layers after adding this module.

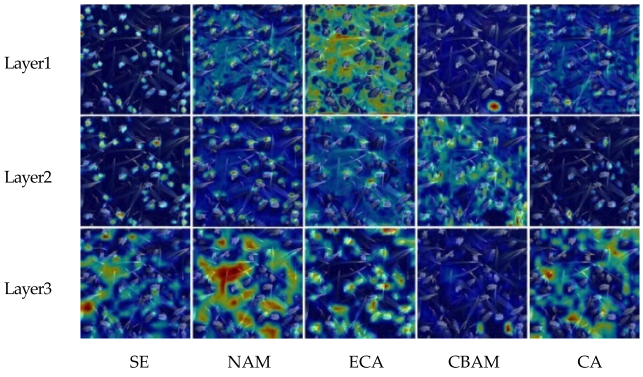

FIGURE 13. Heat Map of models with different attention mechanisms.

From the qualitative analysis of the heatmap image, it can be seen that for the first layer used to detect small targets, CBAM has the worst detection performance. On the contrary, ECA regards the leaves of wheat plants as the targets, perfectly avoiding the targets that should be recognized. Therefore, the model will be greatly affected by the complexity of the background after adding ECA. SE has the best performance, and the other two are also affected by the background to some small targets compared with the detection of medium-sized targets, but CA has missed detection. SE still performs well in medium object detection. In the detection of large targets, ECA performs best, while CBAM is still difficult to detect targets.

From the quantitative analysis of model evaluation results in TABLE 2, CA has the highest accuracy and CBAM has the worst performance. Based on the results of quantitative and qualitative analysis, we choose to add CA after the SPPF

TABLE 2. Results of adding attention mechanism on the model.

Methods	P	R	mAP@0.5	mAP@0.5:0.95	FPS	GFLOPs
YOLOv5	0.910	0.816	0.896	0.462	26.11	107.6
YOLOv5+NMA	0.927	0.848	0.919	0.486	25.13	107.6
YOLOv5+ECA	0.926	0.858	0.924	0.472	24.57	108.1
YOLOv5+CBAM	0.895	0.785	0.867	0.423	25.19	107.6
YOLOv5+SE	0.926	0.852	0.923	0.488	24.94	107.6
YOLOv5+CA	0.923	0.867	0.929	0.491	25.91	107.7

TABLE 3. Comparison of improved models.

Methods	SPConv	CSP	Fusion	CA	mAP@0.5	mAP@0.5:0.95	FPS	GFLOPs
YOLOv5	×	×	×	×	0.896	0.462	26.11	107.6
YOLOV5_A	√	×	×	×	0.922	0.482	12.76	101.0
YOLOv5_B	√	√	×	×	0.918	0.474	26.74	94.3
YOLOv5_C	√	√	√	×	0.925	0.495	25.71	96.8
YOLOv5_D	√	√	√	√	0.938	0.498	27.40	96.9

module. In later experiments, it is also shown that the addition of CA module plays a greater role in improving the model.

2) INFLUENCE OF THE IMPROVED METHOD ON THE MODEL

In this paper, five groups of experiments are designed to compare the influence of all the improved methods on the model. The results of the improved model are shown in TABLE 3, where "√" indicates that the module is used in the model and "×" indicates that the module is not used in the model. Analysis of TABLE 3 shows that YOLOv5_A replaces the vanilla convolution operation in the C3 module in the backbone with the separation-based convolution operation to solve the redundancy problem of the feature mAP, reducing the computational complexity. At the same time, the representative information of high-level features and shallow features are integrated, so that the mAP of the model is greatly improved on the original basis. However, although the complexity of the model is reduced, the inference speed of the model is doubled. The relationship between the complexity of the model and the inference speed is not proportional to each other. Therefore, in order to effectively improve the inference speed of the model, the number of CSP modules in the backbone feature extraction network is reduced in YOLOv5_B, which greatly reduces the complexity of the model and accelerates the inference speed, but the inference accuracy is reduced. In YOLOv5_C, two fusion paths are added to increase the shallow feature information to improve the positioning capability of the model and solve the problem of occluded wheat heads and overlapping of multiple wheat heads. YOLOv5_D introduces the CA mechanism on the basis of YOLOv5_C, which aggregates features from two directions respectively to capture remote dependencies along one spatial direction, while retaining accurate location information along the other spatial direction. This feature strengthens the ability to locate and feature extraction of the model, and greatly improves the accuracy of model inference.

3) COMPARISON BETWEEN THE IMPROVED MODEL AND OTHER MODEL

In order to test the performance and accuracy of the algorithm proposed in this paper, our algorithm is compared with the current mainstream object detection algorithms, and two evaluation indicators of mAP and FPS are used to evaluate and compare each algorithm. The experimental results after comparison are shown in TABLE 4. As can be seen from TABLE 4, our algorithm has the best inference accuracy and inference speed. Faster RCNN has good detection accuracy, but the velocity of model inference is the slowest among all models. YOLOv7 and YOLOX have relatively good performance overall.

We take the images of different growth stages of wheat heads and images with fewer targets as the input of each model and visualize the obtained detection results shown in Fig.15. In the figure, a blue oval shape is used to circle the targets that are not recognized, and a white rectangular box is used to mark the over-predicted detection box. Some models will identify overlapping wheat heads as the same head, which we use purple rectangular boxes for marking. Analyzing the visualization results, it can be seen that the missed detection rate of the model is the highest during the post-flowering period, and only the Faster RCNN model and our algorithm perform well. However, Faster RCNN will over-detect, detecting the background as an object, or obtaining multiple boxes for one object. YOLOv7 [31], YOLOX [32] and SSD [33] identify multiple overlapping targets as one target. On the whole, SSD and YOLOv3_SPP perform the

TABLE 4. Comparison of mainstream target detection models.

Model	mAP@0.5	mAP@0.5:0.95	FPS
Faster RCNN	0.827	0.499	18.52
SSD	0.727	0.466	22.48
YOLOv3_spp	0.763	0.408	27.71
YOLOv5	0.896	0.462	26.11
YOLOv7	0.918	0.477	29.07
YOLOX	0.83	0.55	26.68
ours	0.938	0.498	27.40

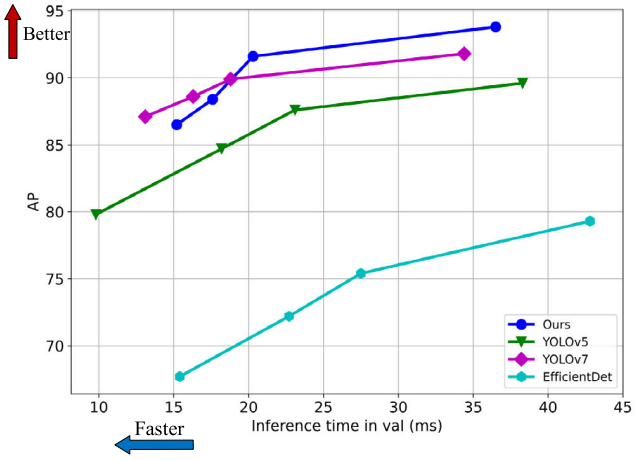

FIGURE 14. Comparison of YOLOv5, YOLOv7 and ours on GWHD_2021 dataset with EfficientDet as the benchmark.

worst. Specifically, the targets with little difference between color and background are missed, and the wheat heads with incomplete appearance displayed at the edge of the image are missed. Our model ensures the correct identification of wheat that adheres to each other. Compared with other models, our model has the lowest rate of missed targets, but there are still some cases that some targets in the edge of image are missed.

In addition, we compare the inference effect of the model under different lighting conditions. This is shown in Fig.16. It can be seen that the model performs poorly in the situation of sunny and backlight, with more missed detection cases than under other lighting conditions. The detection effect is best under normal lighting on cloudy days. Faster RCNN still suffers from over-recognition. In addition, SSD, YOLOv5, and YOLOX all identify the two adhered wheat heads as a single target under the condition of backlight on sunny days. YOLOv5, YOLOv7 and YOLOX do not perform well in recognizing small objects integrated into the background. SSD and YOLOv3_SPP perform the worst in the overall model. YOLOv7 has wrong recognition results for targets under sunny backlight conditions, as shown inside the green rectangular box in the figure, identifying the shadow of wheat plant leaves without objects as wheat ears. In summary, our algorithm has a good adaptability to the detection of wheat head images under various lighting conditions, which indicates that our model has good robustness. Taking into account the influence of different lights, our model also has a good effect on the detection of wheat ears that adhere to each other.

Due to the non-negligible detection accuracy and speed of YOLOv7, we plot the detection accuracy of YOLOv7, YOLOv5 and our model on the GWHD_2021 dataset on different shapes of validation sets. Since the EfficientDet network is used as the benchmark in the YOLO papers, in this article, we follow the tradition of the YOLO papers. The shape of the original validation set was 1024×1024, and then it was adjusted to 512×512, 416×416, 320×320 respectively. The average inference time and average precision of these models on each image were compared, and the results of Fig.14 were obtained. It can be seen that although our model has a good detection effect on images with high resolution, as the resolution of images decreases, our model starts to open a gap with the YOLOv7 model, but it is higher than YOLOv5 in general. After analysis, it is found that YOLOv7 benefits from the fusion of two down-sampling methods (pooling and convolution) during down-sampling period, while our down-sampling only uses a pooling operation. In addition, the RepConv operation of YOLOv7 fuses the convolution and normalization operations into a single convolution operation, which greatly reduces the inference time, and at these two points, is the deficiency of our algorithm.

IV. DISCUSSION

When adding the SPConv module, we used three strategies to determine where it should be located. In the first strategy, the SPConv module was added after each C3 module in backbone, which increased the complexity of the model somewhat, and led to the cost of a 2% reduction in detection accuracy. In the second strategy, all vanilla convolutions in the network are replaced by SPconv, which still greatly reduces the detection accuracy of the model. The third strategy replaces the convolution in the CSP block with SPConv, which improves the detection accuracy and reduces the model complexity. There are two reasons for the accuracy reduction of the first two strategies, one is that some features of the shallow layers are to some extent divided into redundant information proposed, and the other reason is that too many SPConv blocks will filter out part of the useful information.

When determining the number of C3 modules, we tried to reduce the number of C3 modules in backbone from the

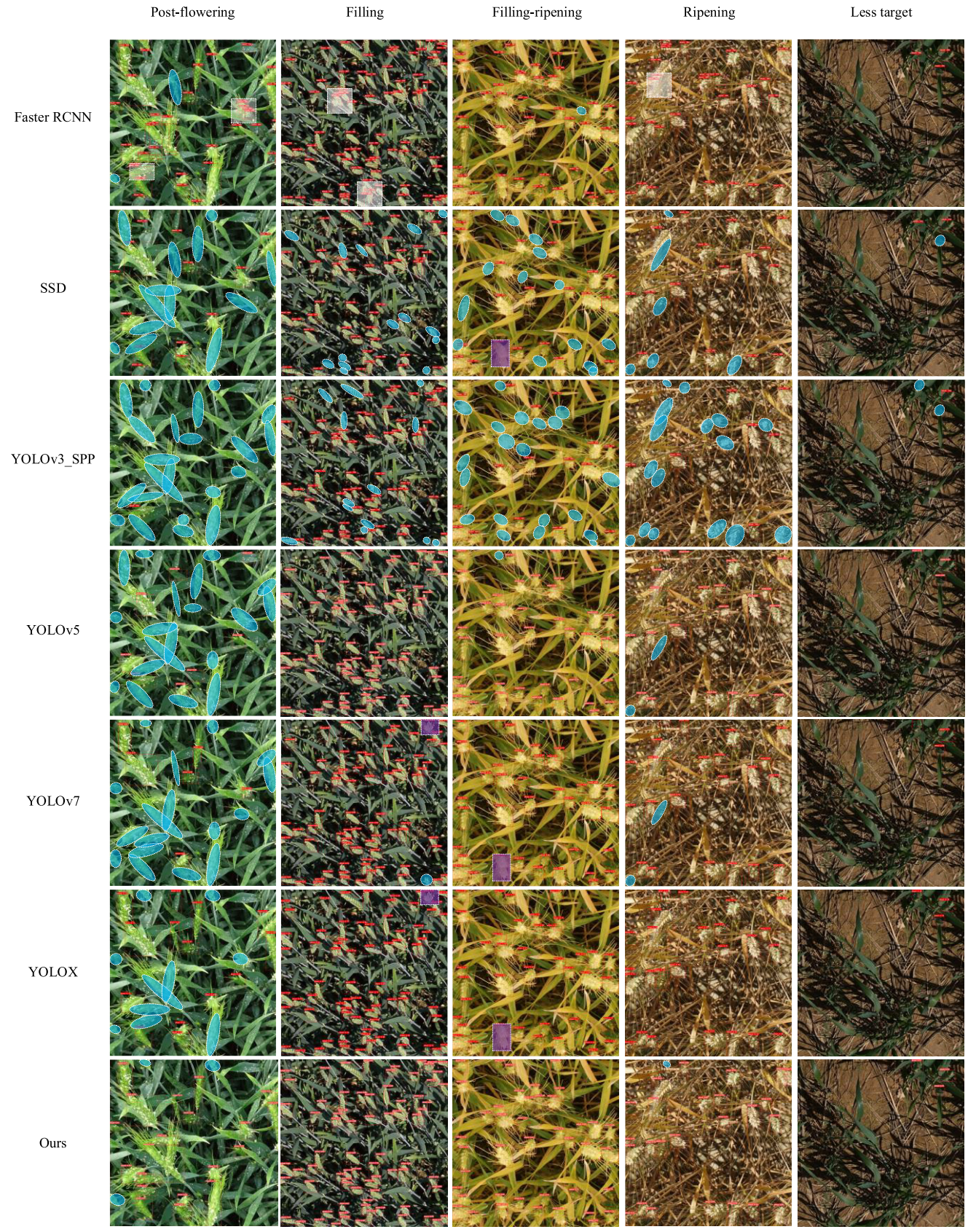

FIGURE 15. Visualization of inference results of different models for wheat ears at different growth periods.

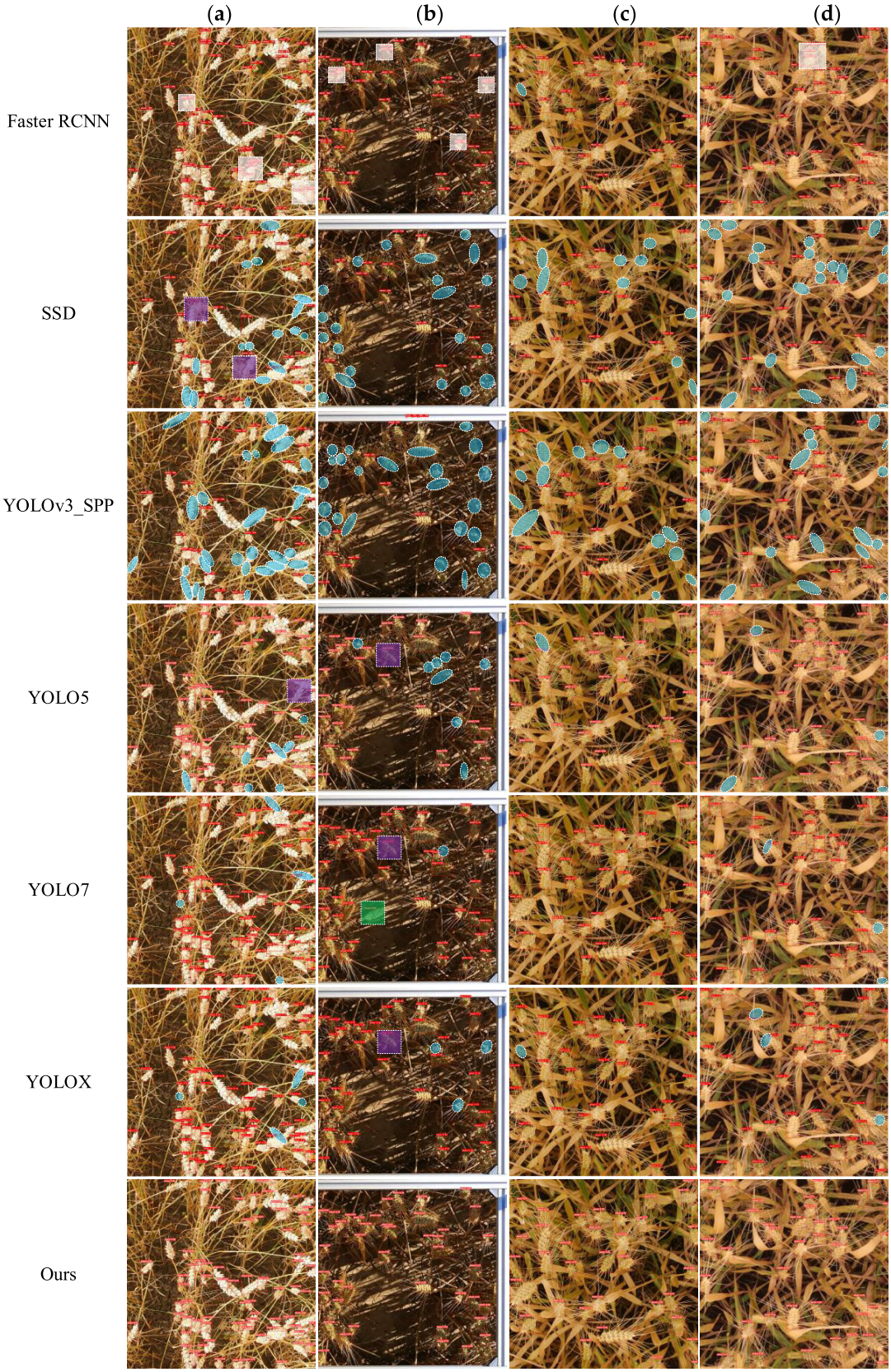

FIGURE 16. Visualization of inference results of different models on images under different lighting conditions. (a)Sunny and direct sunlight condition. (b)sunny and backlight condition. (c)cloudy and direct sunlight condition. (d)cloudy and backlight condition.

original 3, 6, 9, 3 to 1 and 3 respectively. From the experimental results, after changing the number of C3 modules to 1, the computational complexity was greatly reduced, from the original 107 to 72, and the detection accuracy was reduced by 2%. When it comes to 3, mAP remains unchanged when IOU is 0.5, and the computational complexity will be 88, which indicates that the number of C3 modules can be appropriately reduced to reduce the complexity of the model and not greatly reduce the accuracy of model detection, but at the same time, the number of modules should not be too small.

When adding the attention mechanism, we tried five attention mechanism modules separately. The performance of CBAM module is the worst, which indicates that the method like CBAM, which only obtains the recalibrated features by passing the feature map through spatial attention and channel attention in turn, will lose feature information to a certain extent, while CA aggregates features along two directions by embedding spatial location information into channel attention, enhancing the model's perception of direction and sensitivity to position, further improving the model detection accuracy, and the precision of densely distributed target detection.

V. CONCLUSION

The existing YOLOv5 itself has an almost perfect performance, but the result of detection in targets that have similar color with background is not so good, which, we analyze, is related to the filtering of positive and negative samples in classification, and it is also the limitation of the model when dealing with images in low-resolution. From our results, it can be seen that the appropriate fusion of separable convolution in YOLOv5 can make the target extraction network more efficient to extract features, thus, the small error caused by rough screening of positive and negative samples is reduced. Besides, the CA module can enhance the sensitivity of the model to the location, which is of great help for the detection of densely distributed targets. However, our improved model is not perfect for the detection ability of incompletely displayed small objects at the edge position of the image, and it is not very good for the extraction of edge features. In addition, the performance of YOLOv7 and YOLOX is improved compared with the original YOLOv5 model. Therefore, it is possible to refine on more complete models like the two to increase their ability to extract the edge information of the target and adapt to changes in target size, which may lead to unexpected results. But then again, compared with the mainstream algorithms, our algorithm performs the best in the detection of wheat heads with different growth periods, under different lighting conditions, and that adhere to each other, and also has good inference speed. Finally, our model simply realizes the detection, but in practical application, farmers need to estimate the grain yield per unit area according to the detection results. Therefore, future work will mainly focus on counting and establishing the regression model of quantity and yield.

REFERENCES

[1] M. Ghahremani and H. Ghassemian, "Remote-sensing image fusion based on curvelets and ICA," *Int. J. Remote Sens.*, vol. 36, no. 16, pp. 4131–4143, 2015, doi: 10.1080/01431161.2015.1071897.

[2] J. A. Fernandez-Gallego, S. C. Kefauver, N. A. Gutiérrez, M. T. Nieto-Taladriz, and J. L. Araus, "Wheat ear counting in-field conditions: High throughput and low-cost approach using RGB images," *Plant Methods*, vol. 14, no. 1, p. 22, Dec. 2018, doi: 10.1186/s13007-018-0289-4.

[3] C. Zhou, D. Liang, X. Yang, B. Xu, and G. Yang, "Recognition of wheat spike from field based phenotype platform using multi-sensor fusion and improved maximum entropy segmentation algorithms," *Remote Sens.*, vol. 10, no. 2, p. 246, Feb. 2018, doi: 10.3390/rs10020246.

[4] N. Alharbi, J. Zhou, and W. Wang, "Automatic counting of wheat spikes from wheat growth images," in *Proc. 7th Int. Conf. Pattern Recognit. Appl. Methods*, 2018, pp. 1–10.

[5] P. Sadeghi-Tehran, N. Virlet, E. M. Ampe, P. Reyns, and M. J. Hawkesford, "DeepCount: In-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks," (in English), *Frontiers Plant Sci.*, vol. 10, p. 1176, Sep. 2019, doi: 10.3389/fpls.2019.01176.

[6] S. Khan and A. F. Mollah, "Wheat head detection from outdoor wheat field images using YOLOv5,", in *Emerging Technologies in Data Mining and Information Security*. Singapore: Springer, 2023, pp. 535–542.

[7] S. Khaki, N. Safaei, H. Pham, and L. Wang, "WheatNet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting," *Neurocomputing*, vol. 489, pp. 78–89, Jun. 2022.

[8] F. Han and J. Li, "Wheat heads detection via Yolov5 with weighted coordinate attention," in *Proc. 7th Int. Conf. Cloud Comput. Big Data Analytics (ICCCBDA)*, Apr. 2022, pp. 300–306, doi: 10.1109/ICCCBDA55098.2022.9778925.

[9] C. Liu, K. Wang, H. Lu, and Z. Cao, "Dynamic color transform for wheat head detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1278–1283, doi: 10.1109/ICCVW54120.2021.00148.

[10] A. Carlier, S. Dandrifosse, B. Dumont, and B. Mercatoris, "Wheat ear segmentation based on a multisensor system and superpixel classification," *Plant Phenomics*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.34133/2022/9841985.

[11] Y. Dong, Y. Liu, H. Kang, C. Li, P. Liu, and Z. Liu, "Lightweight and efficient neural network with SPSA attention for wheat ear detection," *PeerJ Comput. Sci.*, vol. 8, p. e931, Apr. 2022.

[12] S. Bhagat, M. Kokare, V. Haswani, P. Hambarde, and R. Kamble, "WheatNet-lite: A novel light weight network for wheat head detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1332–1341.

[13] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," *Pattern Recognit.*, vol. 34, no. 12, pp. 2259–2281, 2001, doi: 10.1016/S0031-3203(00)00149-7.

[14] A. Sun, W. Jia, D. Hei, Y. Yang, C. Cheng, J. Li, Z. Wang, and Y. Tang, "Application of concave point matching algorithm in segmenting overlapping coal particles in X-ray images," *Minerals Eng.*, vol. 171, Sep. 2021, Art. no. 107096, doi: 10.1016/j.mineng.2021.107096.

[15] R. Mehrotra, S. Nichani, and N. Ranganathan, "Corner detection," *Pattern Recognit.*, vol. 23, no. 11, pp. 1223–1233, 1990.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[17] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[18] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[20] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.

[22] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13728–13737.

[23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.
[24] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Nov. 2020, doi: 10.1109/TMI.2020.3035253.
[25] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.
[27] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," 2021, *arXiv:2111.12419*.
[28] Q. Zhang, Z. Jiang, Q. Lu, J. Han, Z. Zeng, S. Gao, and A. Men, "Split to be slim: An overlooked redundancy in vanilla convolution," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–7.
[29] E. David, M. Serouart, D. Smith, S. Madec, K. Velumani, S. Liu, X. Wang, F. Pinto, S. Shafiee, I. S. Tahir, and H. Tsujimoto, "Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods," *Plant Phenomics*, vol. 2021, Jan. 2021, Art. no. 9846158, doi: 10.34133/2021/9846158.
[30] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031, doi: 10.1109/ICCV.2019.00612.
[31] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[32] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding Yolo series in 2021," 2021, *arXiv:2107.08430*.
[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 21–37.

RAN SHEN is currently pursuing the master's degree in computer science and technology with the Henan University of Technology. Her research interests include artificial intelligence, computer vision, image processing, and food informatization.

TONG ZHEN is currently a Professor and the Supervisor of doctoral and master's students with the Henan University of Technology. He has published more than 100 papers, three works, hosted and participated in more than ten scientific research projects, and received more than 30 authorized patents. His research interests include computer application and intelligent control. He was a recipient of many awards, such as the Henan Provincial Science and Technology Progress Award and the Henan Provincial Teaching Achievement (First Prize).

ZHIHUI LI has presided over or participated in more than ten national, provincial, and ministerial major science and technology special projects and provincial natural science fund projects. She teaches courses in analog electronic technology, detection and sensing, presided over two teaching and research projects, and won two second prizes in teaching competitions. She has published more than 20 papers and contributed to the publication of two works. She has won the second prize of the Henan Science and Technology Progress Award and two authorized invention patents.