electronics

Article

Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism

Rui Li * and Yanpeng Wu

School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China; yanpengwu6@gmail.com
* Correspondence: lirui@nwnu.edu.cn

Abstract: The detection and counting of wheat ears are essential for crop field management, but the adhesion and obscuration of wheat ears limit detection accuracy, with problems such as false detection, missed detection, and insufficient feature extraction capability. Previous research results have shown that most methods for detecting wheat ears are of two types: colour and texture extracted by machine learning methods or convolutional neural networks. Therefore, we proposed an improved YOLO v5 algorithm based on a shallow feature layer. There are two main core ideas: (1) to increase the perceptual field by adding quadruple down-sampling in the feature pyramid to improve the detection of small targets, and (2) introducing the CBAM attention mechanism into the neural network to solve the problem of gradient disappearance during training. CBAM is a model that includes both spatial and channel attention, and by adding this module, the feature extraction capability of the network can be improved. Finally, to make the model have better generalization ability, we proposed the Mosaic-8 data enhancement method, with adjusted loss function and modified regression formula for the target frame. The experimental results show that the improved algorithm has an mAP of 94.3%, an accuracy of 88.5%, and a recall of 98.1%. Compared with the relevant model, the improvement effect is noticeable. It shows that the model can effectively overcome the noise of the field environment to meet the practical requirements of wheat ear detection and counting.

Keywords: wheat ear; deep learning; CBAM; YOLO v5; detection and counting

check for updates

Citation: Li, R.; Wu, Y. Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism. *Electronics* **2022**, *11*, 1673. https://doi.org/10.3390/electronics11111673

Academic Editor: Byung Cheol Song

Received: 12 April 2022
Accepted: 23 May 2022
Published: 24 May 2022

1. Introduction

Wheat is an important food crop globally and is consumed by almost one-third of the population. Current wheat yield forecasting has become an essential part of agricultural production. It can provide a necessary reference for field management and agricultural decision making [1]. Therefore, the accurate identification and counting of wheat ears are essential for monitoring crop growth, estimating wheat yield, and analysing plant phenotypic characteristics.

The number of wheat ears is mainly gathered by manual field yield prediction, capacity prediction, prediction based on annual scenario [2], and forecast based on remote sensing images [3]. Manual field judgements are mainly empirical, have low accuracy, and are also labour-intensive. Volumetric methods are costly and inefficient in wheat density measurement. Remote sensing is based on satellite images as samples. As these images are distant, they are only suitable for large-scale processing and analysis, resulting in a low accuracy of wheat prediction. Meanwhile, multiple linear regression-based predictions are heavily influenced by variables such as precipitation, making accuracy difficult to guarantee and unsuitable for field yield estimation. Traditional image processing techniques often use moving window methods [4] or superpixel segmentation [5] to harvest the image, extract colour or texture features from the sub-image, then train a classifier and use the classifier to identify the wheat ears and complete the count, or they highlight the wheat ears through image processing methods, such as binarising the image to place the ears after removing the adhesion [6]. In contrast, visual sensors can acquire rich texture and colour information at a lower cost. However, the colour-texture characteristics of wheat affect the detection accuracy. Therefore, detecting and counting wheat ears in the natural environment remains a significant challenge.

In addition, the adhesion and occlusion between wheat ears severely limit the accuracy of wheat ear identification and counting. Several scholars have attempted to successfully detect adhering objects using segmentation techniques such as morphology [7], concave point matching [8], and watershed algorithms [9]. However, the morphological variation of wheat ears in images is considerable. Therefore, morphology-based segmentation cannot necessarily be used to identify wheat ears in the adherence zone, resulting in missed and false detections. In addition, the notch matching algorithm cannot detect objects with sharper edges. However, as we all know, the advantages of wheat ears are not smooth, and burr edges are susceptible and delicate and are particularly prone to blend in with complex backgrounds, making it difficult to obtain smooth edges of wheat ear images, even if their binary images undergo a series of erosion and expansion operations. The watershed algorithm requires the calculation of local extremes. However, there are more local extremes since the wheat ears have a detailed texture. Therefore, a watershed algorithm for detecting wheat ears would lead to over-segmentation. Thus, the problem of how to accurately count mutually-occluded wheat ears remains an urgent one.

With the development of image processing techniques, previous studies [10] have shown that classifiers for wheat ear detection were built using machine learning methods, which led to wheat ear detection and counting. Xu et al. used the k-means algorithm to segment wheat ears to achieve recognition [11]. Although the recognition of wheat ears was performed based on machine learning methods, most of them still required a priori knowledge to artificially set up image features, which led to insufficient recognition accuracy in field environments with noise disturbances such as uneven lighting and complex backgrounds. At the same time, it is difficult to detect and count sheaves of wheat in different scenarios based on traditional machine learning methods due to the lack of generalisation ability of the models.

In recent years, deep learning has achieved impressive results in many fields. Significant progress has been made in target detection technology, one of the core problems in computer vision. Target detection uses image processing, deep learning, and other techniques to classify and locate target objects from an image or video, determine whether the input image contains target information, and select the target location and category. Deep learning-based target detection algorithms are mainly divided into two categories. One is the two-stage detection algorithm represented by R-CNN (region-CNN) [12] and Fast-RCNN (regions with CNN features) [13]. Based on feature extraction, firstly, many candidate regions are generated, and then they are classified and regressed. The other is a single-stage detection algorithm represented by an SSD (Single Shot MultiBox Detector) [14] and the YOLO (you only look once) series of algorithms, which performs classification and regression tasks while generating candidate frames. The specific problem of wheat quantity detection is currently being investigated nationally and internationally. For example, HASAN et al. [15] used R-CNN, and MADEC et al. [16] used Faster-RCNN to detect wheat ears. Studies have also used single-stage target detection algorithms, such as counting wheat sheaves, using target detection algorithms such as YOLO [17]. Xiong et al. used context-enhanced local regression networks to detect and measure wheat sheaves [18]. Brilliant results have also been achieved using convolutional neural networks (CNNs) in detecting and counting wheat ears [19–21]. Although convolutional neural networks are a class of feedforward neural networks that include convolutional computation and have a deep structure, they are one of the representative algorithms for deep learning. However, the core of the CNN-based approach to detection is based on the region proposal method, where a sliding window or extracted proposal is first selected to train the network. Then classification is performed in the region proposal. The limitation of this approach is that background regions are often mistaken for specific targets in object recognition. Wheat images captured in a field environment have many distracting factors such as high plant density, multiple overlaps, uneven lighting, and complex backgrounds. Therefore, there are still problems worth exploring and solving for wheat detection based on deep learning.

The YOLO family has spawned many versions as a representative framework for single-stage detection. YOLO is a high-performance general-purpose target detection model. YOLO v1 [22] uses a single-stage detection algorithm for the two tasks of locating a target and classifying the target object. Subsequently, YOLO v2 [23] improved in three areas: more accurate predictions, faster speed, and more objects recognised than the v1 version. YOLO v3 [24] accelerated the implementation of object detection by introducing multi-scale prediction, core network optimisation, and loss function improvements. YOLO v4 [25] presented an efficient and fast object detection model that significantly reduced the computational number of parameters, making it easier to deploy on general-purpose and hardware devices. Compared with YOLO v4, YOLO v5 has a smaller and more flexible structure, faster image inference, and is closer to natural production and life. However, the actual environment in which wheat is grown is complex. The main problems are (1) severe object occlusion, with wheat ears obscuring and overlapping each other; (2) dense objects, with multiple piles of wheat ears challenging to distinguish; (3) small target objects, with the whole target taking up a smaller proportion of the whole image; (4) complex backgrounds, increasing the difficulty of feature target extraction, and these practical factors undoubtedly affect the detection accuracy of wheat. We also found that YOLO v5 also suffers from insufficient bounding box localisation and has difficulty distinguishing between overlapping detection objects, especially objects such as wheat ears that are heavily occluded. However, the presence of an attention mechanism can effectively solve these problems. When processing information, the attention module resembles the human visual attention mechanism by scanning the global image to obtain the target area that needs to be focused on and then devoting more attention resources to this area to obtain more detailed information related to the target while filtering out the secondary data to improve the model effect. The Convolutional Block Attention Module (CBAM) [26] was integrated into the convolutional module of YOLO v5 to implement the learning of target features and location features in the channel dimension and global spatial dimension, respectively. One of the difficulties for small target detection is that using multi-level convolution operations may lead to small targets, with small pixel occupancy being lost in the process. To solve this problem, we proposed to add a quadruple downsampling layer to the original YOLO v5 feature pyramid to improve the semantic information of small targets and, thus, make the model's prediction more accurate. To the best of our knowledge, there are few reports on the detection and counting of wheat ears by YOLO v5 models using the improved method described above.

Therefore, this study rationalises a novel and simple method based on YOLO v5 to detect the number of wheat ears. Firstly, the real wheat images were pre-processed considering their quality, and on top of the image enhancement, we proposed an 8-Mosaic data enhancement method inspired by the 4-Mosaic of the original YOLO v5 model. At the same time, data enhancement methods such as varying degrees of brightness conversion, increasing contrast by different multiples, and performing random multi-angle rotation were performed on the dataset to greatly enrich the number of samples for wheat ear recognition in complex backgrounds. Next, the colour and texture features of the wheat ear images were extracted, and parameters defining the subsequent training were established. Next, an improved YOLO v5 neural network model was built in PyTorch. The main improvements were: (1) eliminating background interference by using a two-channel (channel and spatial channel) attention mechanism, and (2) enriching the semantic information of small targets by adding 4-fold downsampling layers and improving the feature pyramid structure to improve the robustness of the model. The GWHDD dataset was then divided into a training set and a test set according to a 9:1 ratio, and input and output matrices were created for training and testing. Finally, the loss function was improved to GIOU [27] to speed up the convergence of the model according to the actual recognition situation.

In China, there have been few problems involving deep learning for wheat counting so far. The network model proposed in this paper provides a new idea and direction for the accuracy of wheat counting. It will facilitate the rapid development of sustainable, green, and automated smart agriculture.

The remainder of the paper is organised as follows: Section 2 presents the proposed method for processing wheat ear images and the specification of the improved process. Section 3 presents the experiments and results. Section 4 offers a discussion of the proposed method. Section 5 concludes the paper.

2. Materials and Methods

2.1. Data Processing

2.1.1. Data Acquisition and Annotation

The data sets were from the Global Wheat Head Detection Dataset (GWHDD) [28], including 3376 RGB images (1024 × 1024 pixels) with a total of 145,665 wheat ears. These wheat images come from different regions, including Europe (France, Switzerland, United Kingdom), North America (Canada), Oceania (Australia), and Asia (Japan, China). The acquired images have great differences, including different varieties, different planting conditions, and different image acquisition methods. Therefore, the wheat samples from GWHDD are diverse and typical. Part of the wheat image is shown in Figure 1. The relevant dimensions of the dataset are shown in Table 1.

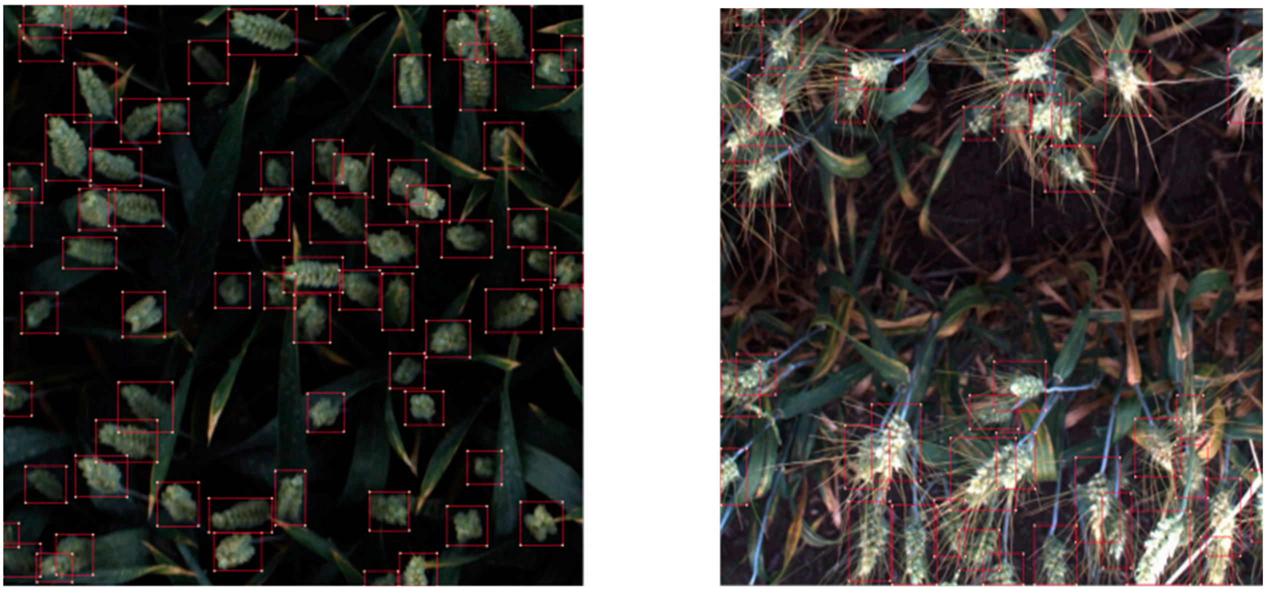


Figure 1. Samples of the wheat ears from the datasets.

Table 1. Wheat related data.

Title 1	Image Number	Width	Height	Wheat Frame	Coordinates Source
0	b6ab77fd7	1024	1024	[834.0, 222.0, 56.0, 360]	usask_1
1	b6ab77fd7	1024	1024	[226.0, 548.0, 130.0, 580]	usask_1
2	b6ab77fd7	1024	1024	[377.0, 504.0, 74.0, 160.0]	usask_1
3	b6ab77fd7	1024	1024	[834.0, 95.0, 109.0, 107.0]	usask_1

In Table 1: width, height (the width and height of the images in the dataset); wheat frame coordinates (the coordinates of the center point of the wheat bounding box in the image, as well as the length and width of the bounding box); source location (the source location of the wheat in the picture).

Based on the data in Table 1, combined with specific wheat pictures, it is clear that the difficulties in detecting wheat ears are as follows: (1) dense wheat plants often overlap; (2) wind can blur photographs; and (3) appearance can vary depending on maturity, colour, genotype, and head orientation, and the geographical source distribution is critical in terms of appearance.

The labelled images comprising the GWHD dataset came from datasets collected between 2016 and 2019 by nine institutions at ten different locations covering genotypes from Europe, North America, Australia, and Asia. These individual datasets are called "sub-datasets." They were acquired over experiments following different growing practices, with row spacing varying from 12.5 cm (ETHZ_1) to 30.5 cm (USask_1). They include normal sowing density (Arvalis_1, Arvalis_2, Arvalis_3, INRAE_1, part of NAU_1) and high sowing density (RRes_1, ETHZ_1, part of NAU_1). The GWHD dataset covers a range of pedoclimatic conditions, including very productive context such as the loamy soil of the Picardy area in France (Arvalis_3), silt-clay soil in mountainous conditions such as the Swiss Plateau (ETHZ_1) or Alpes de Haute Provence (Arvalis_1, Arvalis_2). The geographical source distribution is shown in Figure 2.

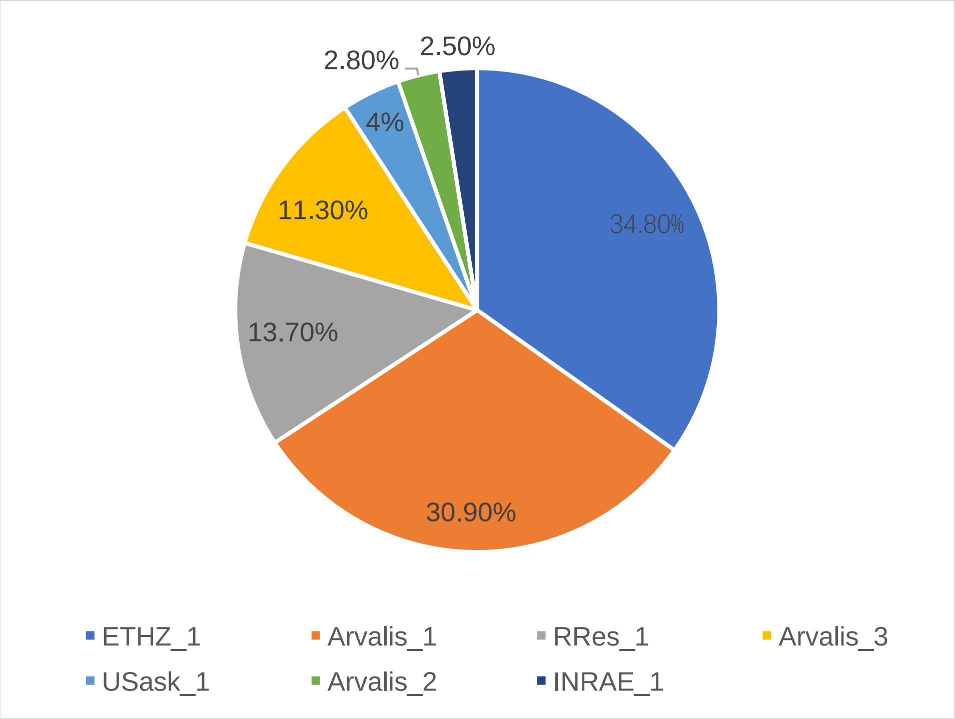


Figure 2. Wheat source distribution map.

Through crawling and other technical means, we finally obtained 3358 photos of wheat ears with different growth environments, different shooting angles, and different maturity stages. The label software was used to label them into the VOC dataset format. At the same time, we manually re-checked the labelled dataset (as shown in Figure 3) to make it as accurate as possible.

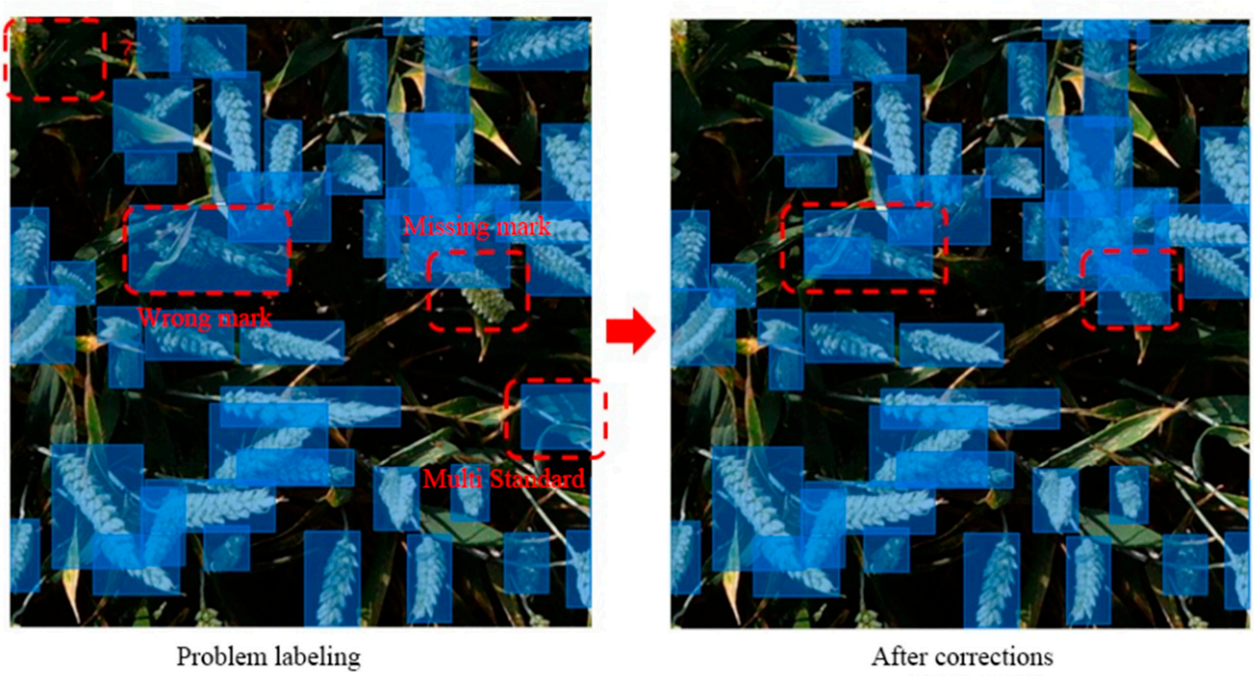


Figure 3. Part of the datasets corrected.

2.1.2. Data Augmentation

To obtain a well-performing neural network model, a large amount of data is often required to support it. However, acquiring new information is often costly in time and labour, so data augmentation is necessary. With data augmentation techniques, the data can be enhanced by using computer-generated data and increasing the amount of data, such as scaling, panning, rotating, colour transformations, and other methods. The advantage of data augmentation is that the number of training samples can be increased, and appropriate noisy data can be added, thus improving the model's generalisation and using the most basic data enhancement method in YOLO v5. The main idea was to cut and scale four pictures randomly and then randomly arrange and stitch them into one image to enrich the dataset and increase small samples. The goal was to improve the training speed of the network. The process of the Mosaic data enhancement is shown in Figure 4.

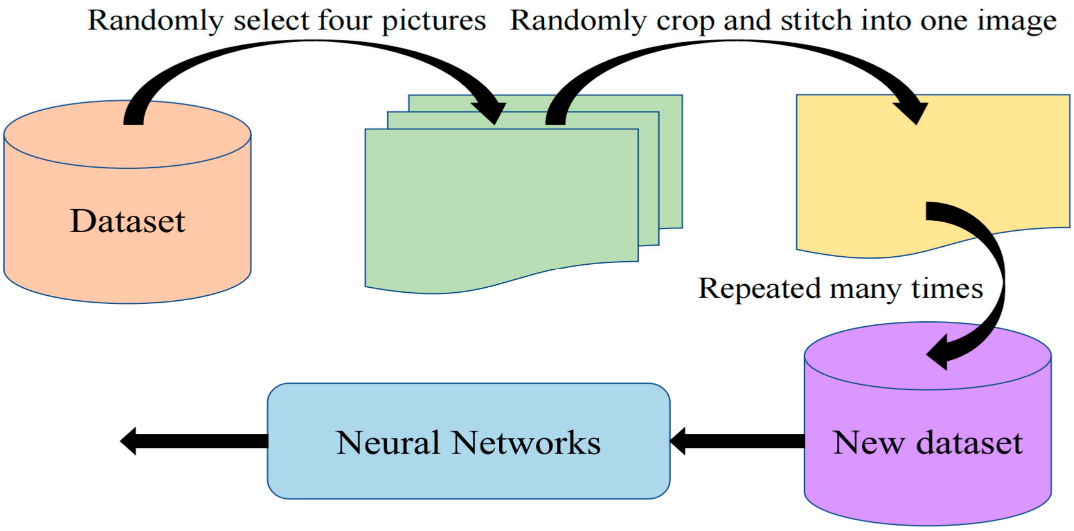

Randomly select four pictures
Randomly crop and stitch into one image
Dataset
Repeated many times
New dataset
Neural Networks

Figure 4. Mosaic data augmentation process.

Inspired by the Mosaic idea, this paper used Mosaic-8, i.e., eight images were randomly cropped, randomly arranged, and randomly scaled, and then they were combined into one image to increase the amount of data in the sample. At the same time, some random noise was reasonably introduced to improve the network model's ability to recognise small target samples in images and enhance the model's generalisation ability, the details of which are shown in Figure 5.

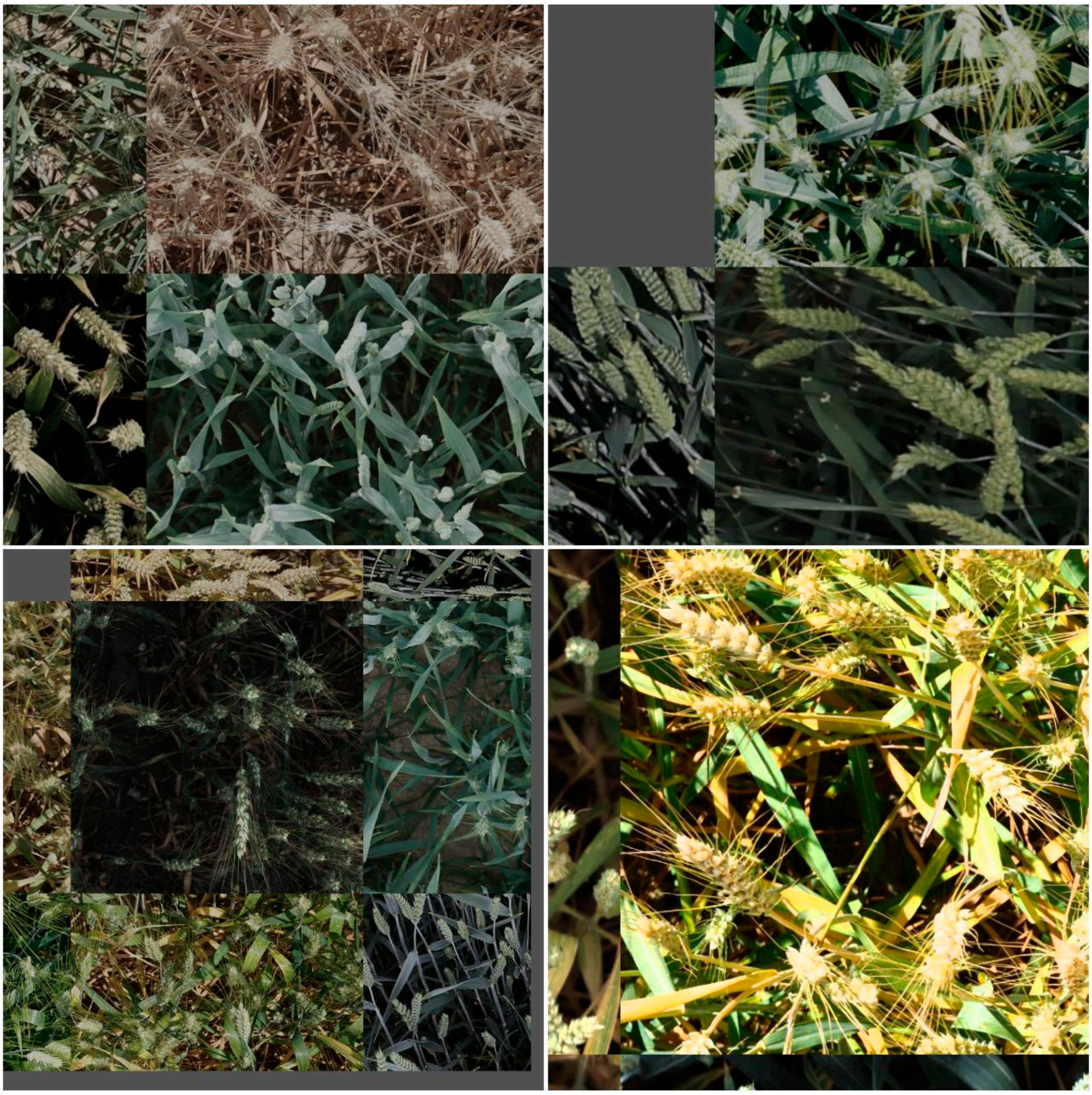


Figure 5. Schematic diagram of part of the Mosaic-8 data augmentation.

In order to improve the robustness of the detection model, a variety of methods were adopted for data enhancement. The specific operations were as follows: (1) perform different levels of brightness conversion, and the brightness of the image was increased by 1.3 times and decreased by 0.7 times, respectively, so that the wheat target detection model was not affected by the diversity of light in the field environment; (2) increase the contrast of the wheat image by 1.2 times and weaken it by 0.8 times, so that the sharpness, gray level, and texture details of the wheat image could be better expressed; (3) perform random multi-angle rotation, such as 90°, 270°, horizontal flip, mirror flip, etc.

2.2. The Improved Network Model

2.2.1. Feature Extractor

In the most primitive YOLO v5 backbone network, three different feature layers are used to detect objects of various sizes. The images enter this detection network and are sequentially passed through the 8× downsampling, 16× downsampling and 32× downsampling layers to obtain three additional sized feature images, fed into the feature fusion network for target recognition. Through the idea of a feature pyramid network [29], we found that, although the feature pictures obtained after deep convolution are rich in semantic information, some target location information will be lost during multiple convolutions, which is not conducive to small target detection. Although the semantic information of the feature image obtained after shallow convolution is not rich enough, the target's location information is more accurate.

In a complex environment, most wheat ears are detected with targets representing only a tiny part of the whole image. Therefore, in this paper, a 4× downsampling layer was added to the YOLO v5 backbone network to improve the location information for small target detection, and the network structure is shown in Figure 6.

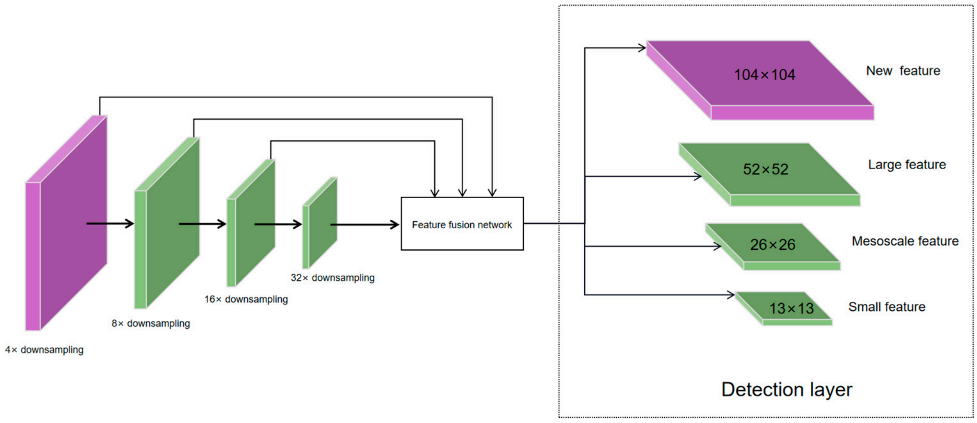


Figure 6. The improved feature extraction model.

After 4× downsampling, the original image is fed into a feature fusion network to obtain a feature map of the new dimension. This feature map has a small field of perception and relatively affluent position information, which improves the detection of small objects.

In convolutional neural networks, the feature pictures we obtained through different convolutional layers contain different target features. The feature images obtained by shallow convolution have high resolution and relatively rich target location information, but the semantic information is not obvious, while the feature images obtained by deep convolution have low resolution and richer semantic information, but the lost target location information is also more abundant. Therefore, the shallow convolution layer can discriminate simple objects, while the deep convolution layer can discriminate complex objects. The fusion of information between the shallow and deep convolutional layers is more beneficial for object detection, which is the principle of feature fusion networks. As we can see in Figure 7, a feature pyramid network is combined with a path aggregation network [30], with the feature pyramid network conveying deep semantic features from top to bottom and the path aggregation network conveying the location information of the target from bottom to top. By fusing top-down and bottom-up feature information, the model can learn features better and improve the accuracy of the model for small target detection.

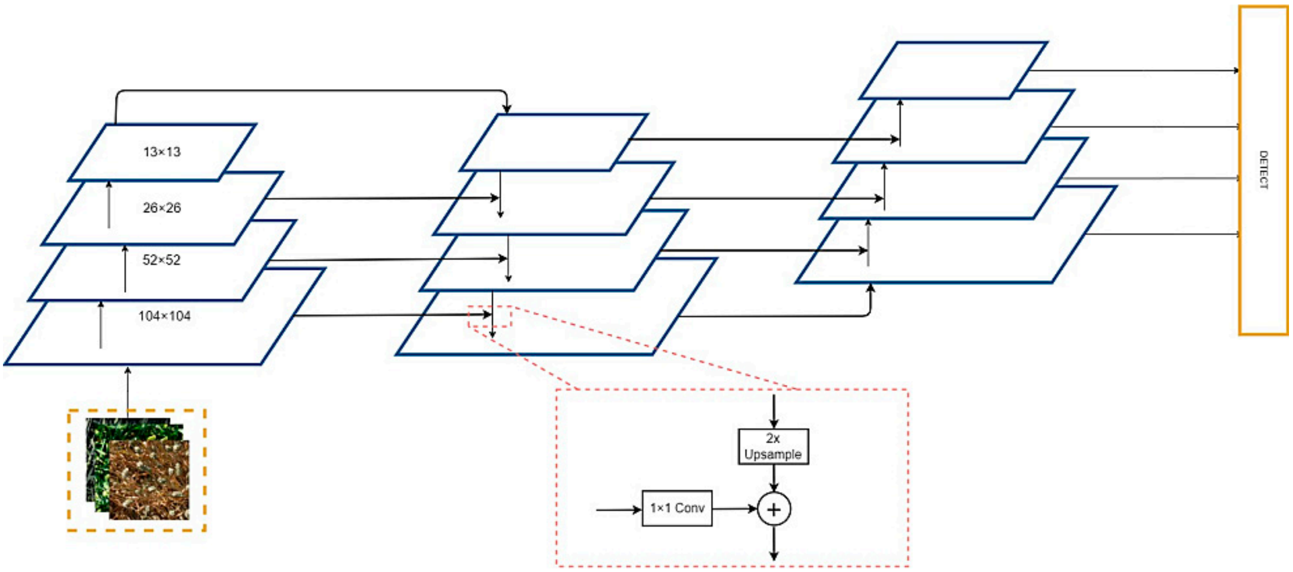


Figure 7. The improved feature fusion network.

2.2.2. Introduce Convolutional Attention Module

Based on the YOLO v5 framework, the convolutional block attention module CBAM (Convolutional Block Attention Network) was introduced. CBAM consists of a channel and spatial attention module, as shown in Figure 8. For input feature F, the global information of each feature channel was obtained by the global average pooling and maximum pooling operations, and then the future channel attention vector was obtained through two fully connected layers, FC1 and FC2, which were used to weigh the input feature F channel by channel to obtain the feature F′.

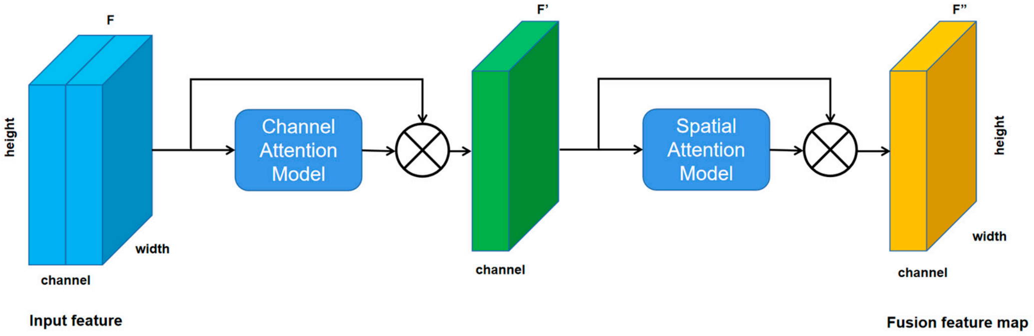


Figure 8. Structure illustration of the channel and spatial attention module.

For the spatial attention module, the feature was subjected to the maximum pooling operation to obtain the feature map, which was used to calculate the spatial information of the feature map. In addition, the feature F′ was input into a 3 × 3 convolutional layer and output by the sigmoid function to obtain the spatial attention map, which was used to activate the feature F′ to obtain the fusion feature F″.

Leaves or other wheat ears usually concealed the information for detecting wheat ear characteristics. Therefore, the channel attention module could enhance the feature expression of the occluded target related to the current task.

2.2.3. Introduce Target Box Regression

Target frame regression aims to find a mapping relationship. The mapping of a candidate target frame (Region Proposal) is infinitely close to that of an actual target frame (Ground Truth). The prediction of the basic structure is made by regressing the relative position of the frame relative to the upper left corner of a grid. The relationship between the a priori frame and the predicted structure is shown in Figure 9, where the dashed frame represents the a priori frame, and the solid frame represents the expected frame. The prediction frame is obtained by translating and scaling the a priori frame. According to the feature map size, the original image is divided into S × S grid cells. Each grid cell will predict three prediction frames, each containing four coordinate information and one confidence level information. When the centre of a target in a natural frame falls in a grid, that grid predicts the target.

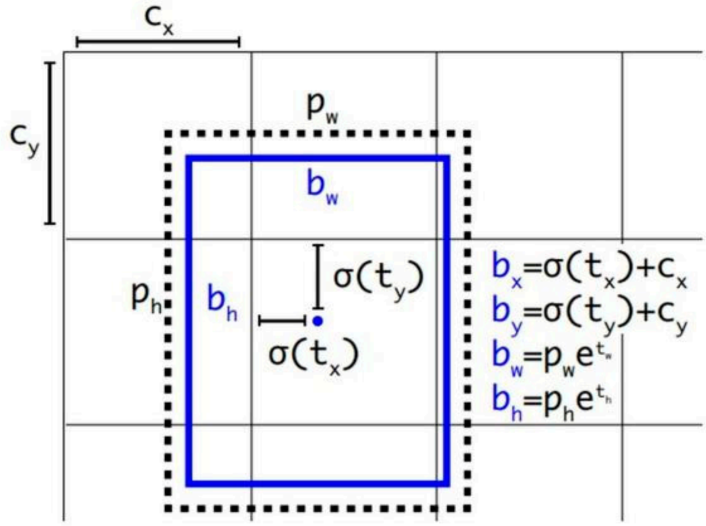


Figure 9. Goal box regression schematic.

The predicted coordinates of the target frame are shown in Equations (1)–(4):

$$b_x = 2\sigma(t_x) - 0.5 + c_x \tag{1}$$

$$b_y = 2\sigma(t_y) - 0.5 + c_y \tag{2}$$

$$b_w = p_w(2\sigma(t_w))^2 \tag{3}$$

$$b_h = p_h(2\sigma(t_h))^2 \tag{4}$$

The network model prediction yields four offsets, t_x, t_y, t_w, t_h, σ, which represent the Sigmoid activation function used to map the network prediction values, t_x, t_y, t_w, t_h to between [0, 1], c_x, c_y is the offset in the cell grid relative to the top left corner of the image, and p_w, p_h is the a priori box width and height. The final centre coordinates b_x, b_y, and width b_w and height b_h of the predicted target box are obtained by the above equation. $\sigma(t_o)$ is the confidence level of the predicted frame, obtained by multiplying the probability of the predicted frame and the IoU value of the predicted frame with the real frame. A threshold is set for $\sigma(t_o)$ to filter out the prediction frames with low confidence, and then the final prediction frame is obtained by using Non-Maximum Suppression (NMS) [31] for the remaining prediction frames.

2.2.4. Loss Function

The original YOLO v5 loss function is shown in Equation (5), consisting of locality loss, confidence loss, and category loss. The confidence loss and category loss are calculated using the binary cross-entropy loss function, as shown in Equations (7) and (8), where K denotes the final output of the network divided into $K \times K$ lattices, M denotes the number of anchor frames corresponding to each lattice, I_{ij}^{obj} denotes anchor frames with targets,

I_{ij}^{noobj} denotes anchor frames without marks, and λ_{noobj} denotes the confidence loss weight coefficient of anchor frames without effects.

$$Loss_{object} = Loss_{loc} + Loss_{conf} + Loss_{class} \tag{5}$$

$$Loss_{loc} = 1 - GIoU \tag{6}$$

$$\begin{aligned} Loss_{conf} = &- \sum_{i=0}^{K \times K} I_{ij}^{obj} [\hat{C}_i^j log(C_i^j) + (1 - \hat{C}_i^j) log(1 - C_i^j)] \\ &- \lambda_{noobj} \sum_{i=0}^{K \times K} \sum_{j=0}^{M} I_{ij}^{noobj} [\hat{C}_i^j log(C_i^j) + (1 - \hat{C}_i^j) log(1 - C_i^j)] \end{aligned} \tag{7}$$

$$Loss_{class} = - \sum_{i=0}^{K \times K} I_{ij}^{obj} \sum_{c \in classes} [\hat{P}_i^j log(P_i^j) + (1 - \hat{P}_i^j) log(1 - P_i^j)] \tag{8}$$

$$GIoU = IoU - \frac{|C - GT \cup P|}{|C|} = \frac{|P \cap GT|}{|P \cup GT|} - \frac{|C - GT \cup P|}{|C|} \tag{9}$$

As shown in Figure 10, the black box is the actual box, denoted GT, the green box is the prediction box, denoted P, and the grey box is the smallest box that wraps both the basic package and the prediction box, marked C, where c is the length of the diagonal of the grey box, and d is the length of the centroid of the actual container and the centroid of the prediction box. The $GIoU$ is calculated as shown in Equation (9).

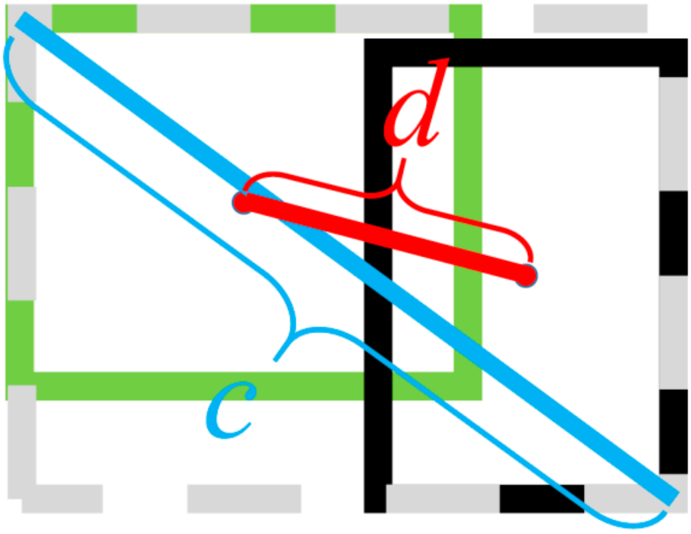


Figure 10. Prediction box P and True box GT.

Unlike the original IoU, GIoU focuses not only on the overlap area between the actual frame and the predicted frame, but also on other non-overlapping areas, so you can better reflect the overlap between the two compared to the original IoU.

On the most miniature feature map, it should be used to detect large targets because its perceptual field is the largest, so large-scale feature maps should be applied to small size a priori frames, and small size feature maps should be applied to large scale a priori frames for the regression of prediction frames. In this paper, a four-scale detection structure was used, and the correspondence between the size of the feature map and the size of the a priori frame for each of the four scales is shown in Table 2.

Table 2. The relationship between the size of the feature map and the size of the a priori box.

Characteristic Map Size	Characteristic Map Size		
13 × 13	[116, 90]	[156, 198]	[373, 326]
26 × 26	[30, 61]	[62, 45]	[59, 119]
52 × 52	[10, 13]	[16, 30]	[33, 23]
104 × 104	[5, 6]	[8, 14]	[15, 11]

2.3. An Improved Wheat Ear Detection Counting Model Based on YOLO v5

Finally, we improved the overall network structure based on the YOLO v5 network, as shown in Figure 11. The network enhances the mining of small target features by adding a feature map with one-fourth of the input image size, using multi-scale feedback to introduce global contextual information to improve the recognition of small targets in images. The CBAM attention module was also introduced to extract helpful location information. The loss function uses GIOU to better describe the regression of target frames in terms of overlap area, centroid distance, and aspect ratio. Based on the original YOLO v5 using Mosaic-8 data enhancement, the formula of target frame regression was modified to improve the convergence accuracy of the model.

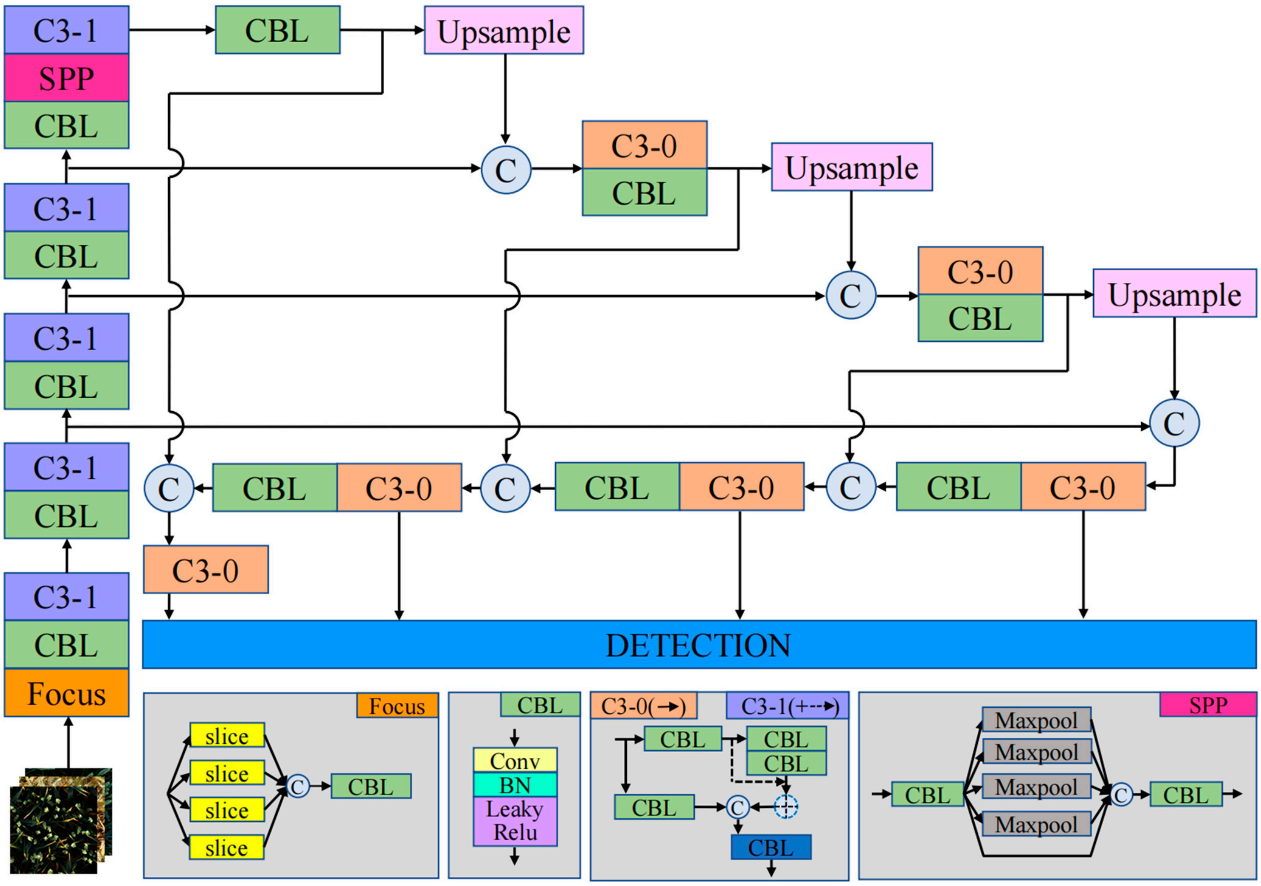


Figure 11. The final modified network structure.

The overall detection process is as follows: first, the training and validation sets were constructed using the manually labelled wheat dataset. Secondly, the training set of wheat images was used to fine-tune the model based on the migration learning approach. Again, the model was tuned, optimised, and validated using the validation set. Finally, the model was tested using the test set, and wheat ear detection and counting results were generated. Among them, the wheatsheaf recognition results are presented in the form of a bounding box. Counting wheat sheaves is based on wheatsheaf recognition, and the results are shown in bounding boxes and statistical ordinal numbers.

2.4. Evaluation of the Model Performance

To test the effectiveness of the model and to verify the transfer performance of the attention information on the model, intersection over union (IOU) was used to evaluate the accuracy of the model according to the coincidence rate of the output box and the label box. Setting a different IOU threshold will result in different numbers of detection frames. Among them, a high threshold results in a small number of detection frames, and a low threshold results in a large number of detection frames. When the detected wheat ear target is small, if a larger threshold is set, the detection of the wheat ear may be missed. Therefore, the threshold value was 0.5 in this study.

In addition, this paper introduced precision (P), which is precision rate, recall rate (R), and mean average precision (mAP), to evaluate the performance of the detection model.

$$\mathrm{P}\ = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$\mathrm{R}\ = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

$$\mathrm{F1}\ - \mathrm{score}\ = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{12}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})d\mathrm{R} \tag{13}$$

$$\mathrm{mAP}\ = \frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i}}{\mathrm{N}} \tag{14}$$

Among them, true positives (TP) mean that both the detection result and the true value are wheat ears, that is, the number of wheat ears was detected correctly. False positives (FP) indicate that the detection result is wheat ears, and the true value is the background, that is, the number of wheat ears counted incorrectly. False negatives (FN) mean that the detection result is the background, and the true value is the wheat ears, that is, the number of wheat ears that are not counted.

"TP + FP" refers to the total number of wheat ears detected, and "TP + FN" refers to the total number of wheat ears in an image. F1-score was used to evaluate the method's performance by balancing the weights of precision and recall.

TP, FP, and FN in precision (Formula (10)) and recall (Formula (11)) represent the number of correctly detected, incorrectly detected, and missing frames. Where the AP value refers to the area of the P-R curve, this paper used the interpolation method to calculate the integral in Formula (13). In Formula (14), the value of mAP is obtained by averaging all categories of AP. N represents the total number of types detected. The larger the value of mAP in this experiment, the better the algorithm detected and the higher the recognition accuracy. The coefficient of determination (R^2), root mean square error (*RMSE*), and Bias were used as evaluation indicators to measure the counting performance of the model, which are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(m_i - c_i)^2}{\sum_{i=1}^{n}(m_i - \overline{m})^2} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(m_i - c_i)^2} \tag{16}$$

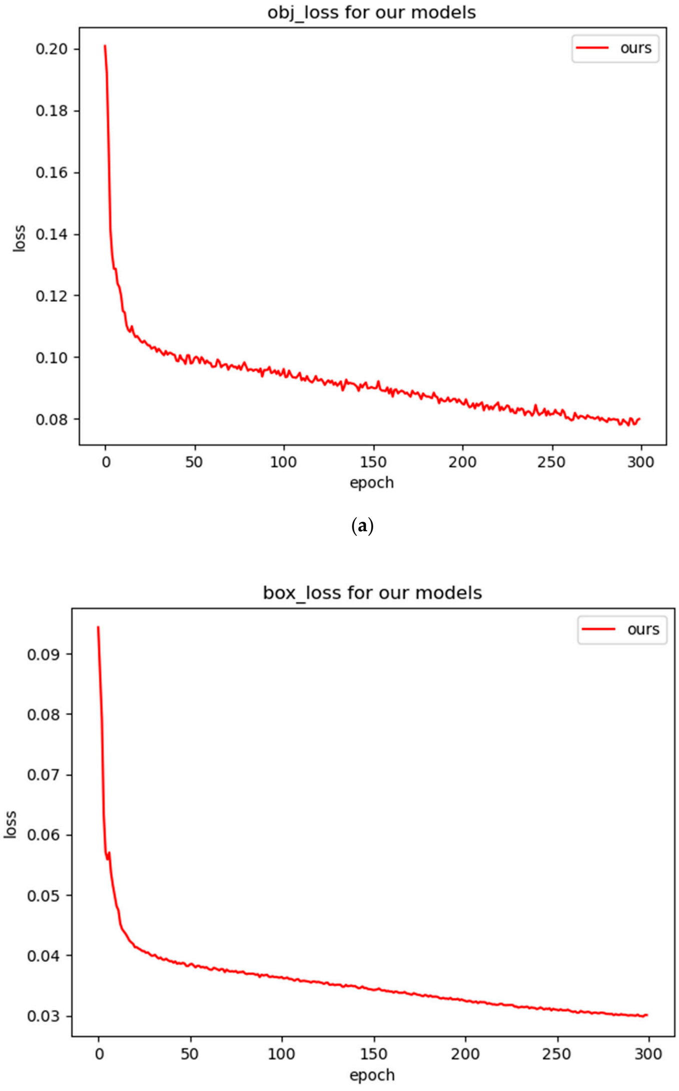
$$Bias = \frac{1}{n}\sum\nolimits_{i=1}^{n}(m_i - c_i) \tag{17}$$

where n represents the number of wheat ear pictures, mi, c_i represent the number of wheat ears manually labeled and counted by the model in the i-th image, and $\overline{m}$ represents the average number of wheat ears. Bias represents the average number of ears of each image detected and the actual error.

3. Results

3.1. Model Training

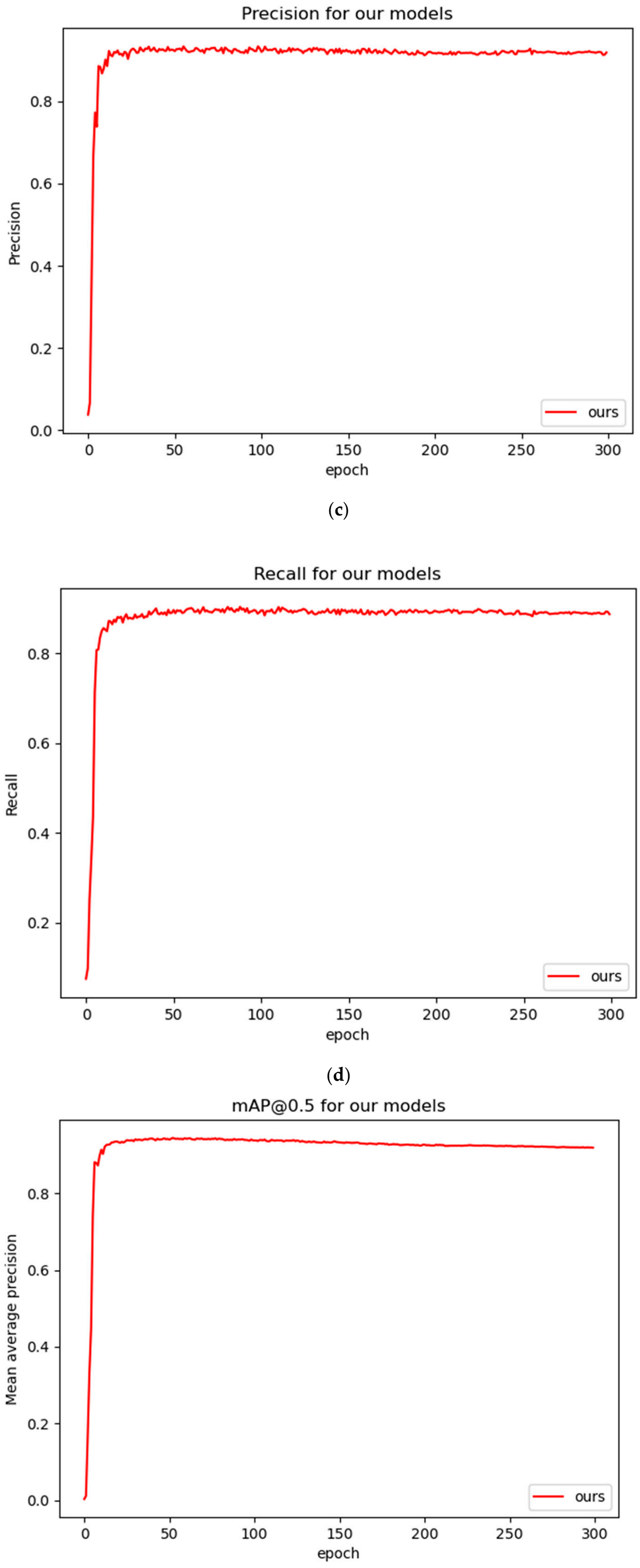
The model training and validation were performed using the training set and validation set in Table 1. The model training parameters were set as follows: learning rate = 0.001, momentum = 0.9, decay = 0.0005, batch size = 16. Adaptive Momentum Estimation (Adam) was used to optimize the training model. The hardware parameters for the experiments were an Intel Xeon Gold 6130H processor and an NVIDIA GeForce GTX 1080Ti GPU, implemented using the Pytorch deep learning framework and Python programming. The CUDA 11.4.0 parallel computing framework and the CUDNN 8.2 deep neural network acceleration library were used in this study, and the total number of iterations in the experiments was 300. Figure 12 shows the model verification accuracy and training loss values obtained in each iteration during the training process.



(**a**)

(**b**)

Figure 12. *Cont.*



(**c**)

(**d**)

(**e**)

Figure 12. Training-process curves. (**a**) Training process: regression loss of bounding box, (**b**) training: confidence loss of bounding box, (**c**) precision, (**d**) recall, and (**e**) mean average precision.

It can be seen from Figure 12 that the training accuracy of the model gradually increased as the number of iterations increased, and the training loss value of the model gradually decreased as the number of iterations increased. The model learning efficiency was high in the initial model training stage, and the training loss curve converged more quickly. After 50 iterations, the loss value decreased slowly. When the iteration reached 200 times, the loss curve was almost flat. As the number of iterations increased, the slope of the training loss curve gradually decreased. Finally, when the number of training iterations reached about 300, the fluctuation trend of the loss value gradually stabilised, and the corresponding accuracy no longer changed.

At the same time, it could be seen from Table 3 that our model was effective in detecting wheat ears, even though the shape, colour, and texture of wheat ears in the image were different. The F1-score was 89.3%, the mAP was 94.3%, the precision was 88.5%, and recall was 98.0%, indicating that our model did not have problems such as over-fitting or under-fitting and gradient disappearance.

Table 3. Evaluation indicators of wheat ear detection.

Data	F1-Score/%	Precision	Recall	mAP
GWHDD	89.3%	88.5%	98.0%	94.3%

It can be seen from Figure 13 that the wheat in the image can be detected by image enhancement, and the trained model can be used for counting wheat ears.

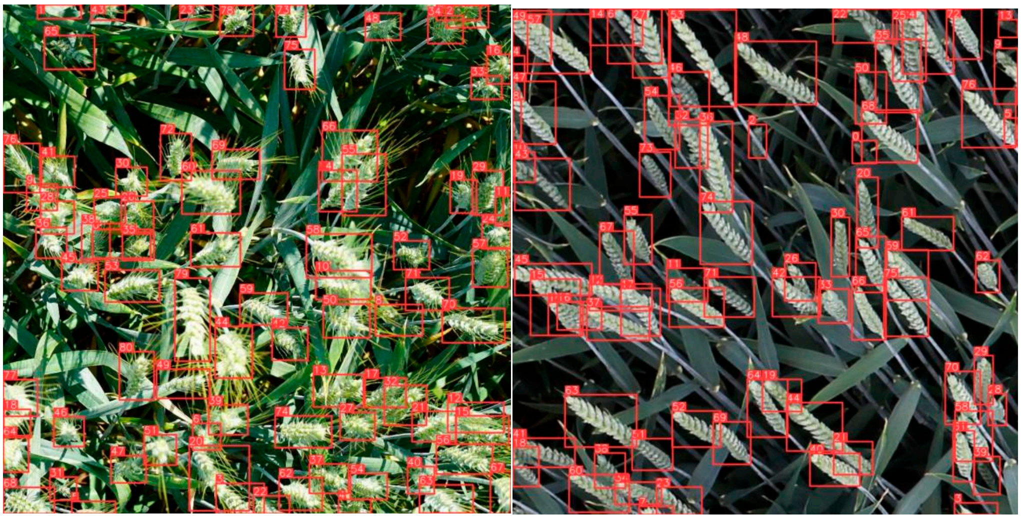


Figure 13. Examples of wheat ear detection results.

3.2. Results of Detecting Wheat Ears

To evaluate the performance of the YOLO v5 model proposed in this study, we used the test set to compare the detection of wheat ear counts by typical convolutional neural networks such as Faster-RCNN, SSD, YOLO v3, YOLO v4, YOLO v5, and some references. The test set results are shown in Table 4, from which we can see that our model had a great improvement compared with other models. If the same series of YOLO v5 is used as the benchmark, then the mAP has increased by nearly three percentage points, which shows that the model has beneficial for detecting wheat ears.

To compare the effect of the proposed method of estimating the number of wheat ears, 20 images were randomly selected from each test set of GWHDD images. Our model was used to detect and count the number of wheat ears. The results are shown in Figure 14. The R^2 of the model was 0.976 for GWHDD, and the RMSE corresponding to the data sets was 3.178. In addition, the results of our method count deviation values that are 2.37.

Table 4. The results of wheat ear detection based on different methods.

Model	Precision	Recall	mAP (%)
Faster-RCNN	47.5%	46.9%	39.2%
SSD	91.2%	36.7%	69.5%
YOLO v3	76.6%	92.2%	88.8%
Reference [32]	76.9%	93.1%	89.5%
YOLO v4	77.8%	93.4%	90.3%
Reference [33]	87.5%	91.0%	93.1%
YOLO v5	88.7%	98.0%	91.9%
Our method	88.5%	98.0%	94.3%

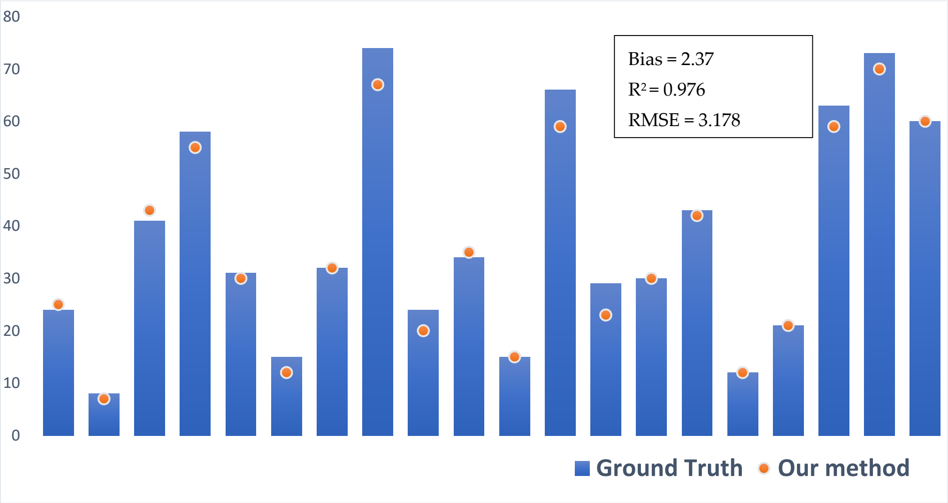
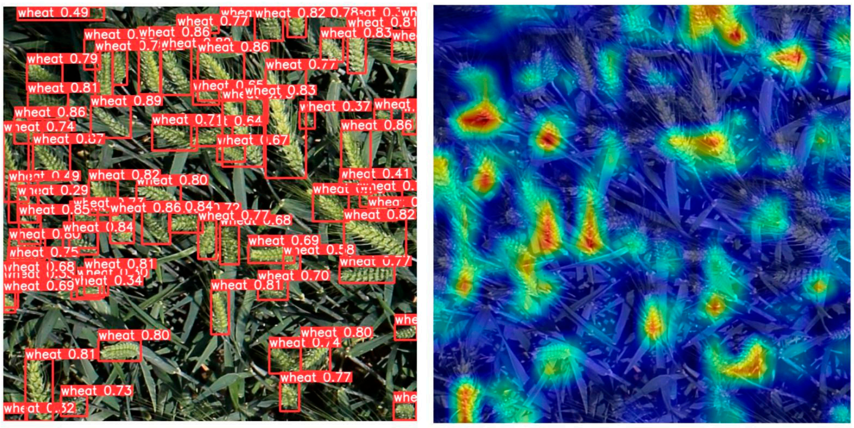


Figure 14. The comparison between the true value and the estimated value of a single image.

Figure 15 shows part of the counting results from the data sets, which were counted based on the proposed method, and the results were compared with the ground truth. It was easy to see that the method we proposed had better robustness, and even if the distribution of wheat ears was relatively concentrated, better counting results could be obtained. At the same time, we also found that the improved YOLO v5 method could detect wheat ears with a complex background, but some of the detection results were higher than the ground truth. A possible reason was that the sample label was insufficient, which led to some wheat leaves being misjudged.



Our method: 74

GT: 67

Figure 15. The comparison between the ground truth and the counting with our method. GT: ground truth.

4. Discussion

Comparison of the Effect of Wheat Ear Detection and Counting under a Complex Background

The detection accuracy of the model under the complex background of the natural environment will be affected to a certain extent. In particular, it was difficult to detect wheat ears when the leaves covered the samples and the wheat ear samples overlapped each other. To test the detection effect of the model proposed in this study under a complex background, 20 images with severe occlusion were selected as data set A and 20 images with slightly occluded wheat ears as data set B. The degree of occlusion was used as a control variable, and our model was used to detect data sets A, B, and A + B, respectively. The detection results are shown in Table 5 and Figure 16. For the detection of lightly obscured wheat ears (data sets A), the F1-score of the model can reach 0.941, and the mAP can reach 97.2%. In a heavily occluded environment with dense targets (data sets B), the model can also achieve an F1 value of 0.923 and an mAP value of 95.8%. The two data sets were mixed into one data set (A + B), and the F1-score and the mAP of the model reached 0.932 and 96.7%. It showed that our model could effectively detect wheat ears in the natural field environment.

Table 5. Evaluation indicators of wheat ear detection.

Test Set	Precision	Recall	F1-Score	mAP (%)
A	0.941	0.936	0.941	97.2%
B	0.948	0.902	0.923	95.8%
A + B	0.955	0.910	0.932	96.7%

A = 20 images with severely occluded wheat ears; B = 20 images with lightly occluded wheat ears.

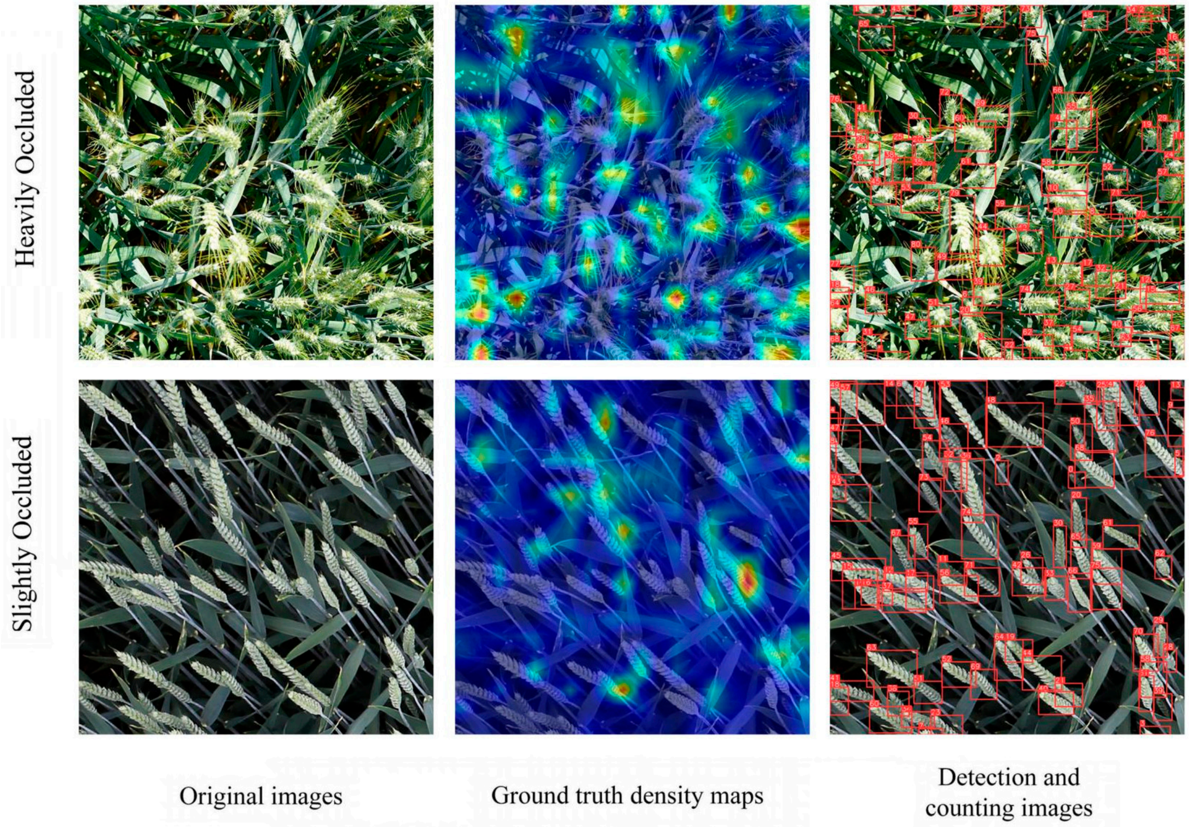


Figure 16. Comparison of the detection results of wheat ears with different degrees of occlusion.

As shown in the density maps in Figure 16, the colours in the maps reflect the size of the density value. The darker the colour, the greater the density value. It could be seen from Figure 16 that the detection and counting of wheat ears can meet the needs of conventional farmland management, regardless of whether the wheat ears are severely occluded or slightly occluded. On the one hand, high-density wheat planting will result in denser wheat ears and severe occlusion. On the other hand, the camera shooting angle will increase the sample occluded by wheat ears [34]. In fact, in fields with dense heat ears, even experts must count the number of wheat ears multiple times to obtain reliable measurement results.

To further verify the detection performance of the algorithm proposed in this study and to explore the effectiveness of each improved method, we designed three ablation experiments based on YOLO v5, each using the same hyperparameters and training techniques. It could be seen from Table 6 that the precision, recall, and mAP of our model were the highest. Among them, the mAP of YOLO v5 + 4× + CBAM was 94.42%, 2.82%, and 0.29% higher than that of the YOLO v5 and YOLO v5 + 4× models, respectively. The results showed that our proposed method improves the model's detection accuracy.

Table 6. Different algorithms for wheat ear detection.

Model	Precision	Recall	mAP (%)
YOLO v5	88.70%	98.01%	91.60%
YOLO v5 + 4×	89.63%	98.04%	94.13%
YOLO v5 + 4× + CBAM	88.52%	98.06%	94.32%

Figure 17 shows the precision-recall curves of three YOLO models on wheat ears. It can be seen from Figure 11 that when the recall of the three models was less than 0.1, the precision remained around 1.0, and the difference was not significant. However, we can see that increasing the shallow feature layer had a greater effect on small targets and mAP increased by 2.72%. Although the mAP only increased by 0.19% after adding CBAM, its corresponding precision was larger than the other two models, indicating that spatial attention could improve the detection performance of the model and could fully reflect the advantages of a spatial attention module.

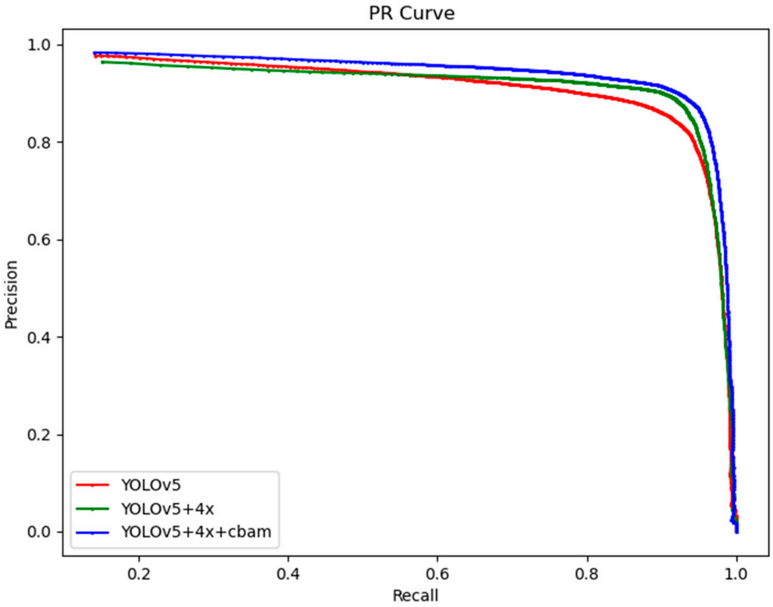


Figure 17. Precision recall curves of the detection results of different algorithms.

5. Conclusions

Wheat counting is an important research area in agricultural yield estimation. Based on the original YOLO v5 algorithm, this paper improved on five aspects: Mosaic data enhancement, feature extraction, loss function, target frame regression, and attention mechanism, which effectively improved the detection accuracy of the YOLO v5 network model for small target objects. Finally, a neural network model for training image recognition was developed. The conclusions can be summarised as follows:

(1) Convolutional neural networks can play an important role in studying wheat counting, using the improved YOLO v5 to detect wheat counts. Its recognition speed is fast, and the recognition rate is high. After pre-processing, feature extraction, and network optimisation, the training recognition rate can reach 94.42%.

(2) The application of Mosaic-8 in the image pre-processing process can accelerate the convergence speed of the model.

(3) The recognition rate of colour and texture features in complex backgrounds can be significantly improved by introducing a shallow feature layer and an attention mechanism.

(4) The method achieves fast and accurate recognition of what counts while also eliminating the need to rely on complex technologies such as remote sensing, which helps to significantly reduce the labour intensity of agricultural practitioners and thus improve work efficiency. In the future, there is a need to increase the number of images collected at different heights to cover the whole process of wheat counting and to increase the number of training samples to improve the recognition rate. In addition, detection methods can be combined with drones to compress the number of model parameters and achieve real-time detection.

The neural network model proposed in this report will shed new light on wheat counting. Due to its unique advantages, crop yield estimation will break new ground. However, fast detection still needs specific hardware configuration. We will continue to optimise our model and use pruning technology to optimise the model. At the same time, we will continue to increase the research on more wheat varieties and increase the scope of application. We believe it can make a great contribution to sustainable, green, and automated agriculture.

Author Contributions: Conceptualization, R.L.; writing—review and validation, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We are grateful to the reviewers for their suggestions and comments, which significantly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Slafer, G.A.; Savin Sadras, V.O. Coarse and fine regulation of wheat yield components in response to genotype and environment. *Field Crop. Res.* **2014**, *157*, 71–83. [CrossRef]
2. Ying, H.; Jian, W.; Mao, R.; Ying, S. Remote Sensing Model for Dynamic Prediction of Maize and Wheat Yield in the United States. *J. Ecol.* **2009**, *28*, 2142–2146.
3. Xiu, Z.; Xiu, L.; Wen, Y.; Huai, Y. Annual prediction model of wheat yield in Rizhao city based on SPSS. *China Agric. Bull.* **2010**, *26*, 295–297.
4. Liu, T.; Sun, C.M.; Wang, L.J.; Song, X.C.; Zhu, X.K.; Guo, W.S. Image processing technology-based counting of wheat ears in large fields. *J. Agric. Mach.* **2014**, *45*, 282–290.
5. Du, Y.; Cai, Y.C.; Tan, C.W.; Li, Z.H.; Yang, G.J.; Feng, H.K.; Han, D. A method for counting the number of wheat spikes in the field based on super pixel segmentation. *Chin. Agric. Sci.* **2019**, *52*, 21–33.
6. Li, Y.; Du, S.; Yao, M.; Yi, Y.; Yang, J.; Ding, Q.; He, R. Field wheat ear counting and yield prediction method based on wheat population images. *J. Agric. Eng.* **2018**, *34*, 193–202.
7. Zhou, C.; Liang, D.; Yang, X.; Xu, B.; Yang, G. Recognition of Wheat Spike from Field Based Phenotype Platform Using Multi-Sensor Fusion and Improved Maximum Entropy Segmentation Algorithms. *Remote Sens.* **2018**, *10*, 246. [CrossRef]
8. Li, Y.; Shiwei, D.; Min, Y.; Yingwu, Y.; Jianfeng, Y.; Qishuo, D.; Ruiyin, H. Method for wheatear counting and yield predicting based on image of wheatear population in field. *Nongye Gongcheng Xuebao/Trans. Chin. Soc. Agric. Eng.* **2018**, *21*, 185–194.
9. Shrestha, B.L.; Kang, Y.M.; Yu, D.; Baik, O.D. A two-camera machine vision approach to separating and identifying laboratory sprouted wheat kernels. *Biosyst. Eng.* **2016**, *147*, 265–273. [CrossRef]
10. Jose, A.F.; Lootens, P.; Irene, B.; Derycke, V.; Kefauver, S.C. Automatic wheat ear counting using machine learning based on RGB UAV imagery. *Plant J.* **2020**, *103*, 1603–1613.

11. Xu, X.; Li, H.; Yin, F.; Xi, L.; Ma, X. Wheat ear counting using k-means clustering segmentation and convolutional neural network. *Plant Methods* **2020**, *16*, 106. [CrossRef] [PubMed]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–28 June 2014; pp. 580–587.
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. *SSD: Single Shot MultiBox Detector*; Springer: Cham, Switzerland, 2016.
15. Hasan, M.M.; Chopin, J.P.; Laga, H.; Miklavcic, S.J. Detection and analysis of wheat spikes using Convolutional Neural Networks. *Plant Methods* **2018**, *14*, 1–13. [CrossRef] [PubMed]
16. Madec, S.; Jin, X.; Lu, H.; De Solan, B.; Liu, S.; Duyme, F.; Baret, F. Ear density estimation from high resolution RGB imagery using deep learning technique. *Agric. For. Meteorol.* **2019**, *264*, 225–234. [CrossRef]
17. Liang, X.; Chen, Q.; Dong, C.; Yang, C. Study on Maize Ear Detection Based on deep learning and unmanned aerial vehicle remote sensing technology. *Fujian J. Agric.* **2020**, *35*, 456–464.
18. Xiong, H.; Cao, Z.; Lu, H.; Madec, S.; Shen, C. TasselNetv2: In-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods* **2019**, *15*, 150. [CrossRef]
19. Yao, J.; Qi, J.; Zhang, J.; Shao, H.; Yang, J.; Li, X. A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5. *Electronics* **2021**, *10*, 1711. [CrossRef]
20. Wang, D.; Fu, Y.; Yang, G.; Yang, X.; Zhang, D. Combined use of FCN and Harris corner detection for counting wheat ears in field conditions. *IEEE Access* **2019**, *7*, 178930–178941. [CrossRef]
21. Sadeghi-Tehran, P.; Virlet, N.; Ampe, E.M.; Reyns, P.; Hawkesford, M.J. DeepCount: In-Field Automatic Quantification of Wheat Spikes Using Simple Linear Iterative Clustering and Deep Convolutional Neural Networks. *Front. Plant Sci.* **2019**, *10*, 1176. [CrossRef]
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
24. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
26. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; p. 11211.
27. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
28. David, E.; Madec, S.; Sadeghi-Tehran, P.; Aasen, H.; Zheng, B.; Liu, S.; Kirchgessner, N.; Ishikawa, G.; Nagasawa, K.; Badhon, M.A.; et al. Global wheat head detection (GWHD) dataset: A large and diverse dataset of high resolution RGB labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics* **2020**, *2020*, 3521852. [CrossRef] [PubMed]
29. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
30. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
31. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), IEEE, Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
32. Wang, Y.; Zhang, Y.; Huang, L.; Zhao, F. Study on wheat ear target detection algorithm based on convolution neural network. *Softw. Eng.* **2021**, *24*, 6–10.
33. Zhu, Y. Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module. *Agronomy* **2021**, *11*, 1202.
34. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance Problems in Object Detection: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3388–3415. [CrossRef] [PubMed]