agronomy

MDPI

Article

Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module

Baohua Yang [1,2,*], Zhiwei Gao [1], Yuan Gao [1] and Yue Zhu [1]

[1] School of Information and Computer, Anhui Agricultural University, Hefei 230036, China; gaozhiwei@ahau.edu.cn (Z.G.); gaoyuan@ahau.edu.cn (Y.G.); zhuyue@ahau.edu.cn (Y.Z.)
[2] Smart Agriculture Research Institute, Anhui Agricultural University, Hefei 230036, China
* Correspondence: ybh@ahau.edu.cn

Abstract: The detection and counting of wheat ears are very important for crop field management, yield estimation, and phenotypic analysis. Previous studies have shown that most methods for detecting wheat ears were based on shallow features such as color and texture extracted by machine learning methods, which have obtained good results. However, due to the lack of robustness of these features, it was difficult for the above-mentioned methods to meet the detection and counting of wheat ears in natural scenes. Other studies have shown that convolutional neural network (CNN) methods could be used to achieve wheat ear detection and counting. However, the adhesion and occlusion of wheat ears limit the accuracy of detection. Therefore, to improve the accuracy of wheat ear detection and counting in the field, an improved YOLOv4 (you only look once v4) with CBAM (convolutional block attention module) including spatial and channel attention model was proposed that could enhance the feature extraction capabilities of the network by adding receptive field modules. In addition, to improve the generalization ability of the model, not only local wheat data (WD), but also two public data sets (WEDD and GWHDD) were used to construct the training set, the validation set, and the test set. The results showed that the model could effectively overcome the noise in the field environment and realize accurate detection and counting of wheat ears with different density distributions. The average accuracy of wheat ear detection was 94%, 96.04%, and 93.11%. Moreover, the wheat ears were counted on 60 wheat images. The results showed that R^2 = 0.8968 for WD, 0.955 for WEDD, and 0.9884 for GWHDD. In short, the CBAM-YOLOv4 model could meet the actual requirements of wheat ear detection and counting, which provided technical support for other high-throughput parameters of the extraction of crops.

Keywords: wheat ear; attention; you only look once (YOLO); detection and counting

Citation: Yang, B.; Gao, Z.; Gao, Y.; Zhu, Y. Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module. *Agronomy* **2021**, *11*, 1202. https://doi.org/10.3390/agronomy11061202

Academic Editors: Saeid Homayouni, Yacine Bouroubi and Karem Chokmani

Received: 2 May 2021
Accepted: 7 June 2021
Published: 12 June 2021

1. Introduction

Wheat is one of the most important food crops and plays an important role in food security. The forecast of wheat yield has become an important part of the agricultural production process, which can provide necessary references for field management and agricultural decision-making [1]. Therefore, accurately identifying and counting wheat ears is of great significance for monitoring crop growth, estimating wheat yield, and analyzing plant phenotypic characteristics.

At present, with the rapid development of machine vision technology, LiDAR [2], heat map [3], and digital images have achieved good detection results in wheat monitoring. In particular, high-resolution images have made the detection of crops easier and more efficient [4]. Among them, image processing and feature extraction have gradually become key technologies for wheat ear recognition, and have made excellent contributions to improving the accuracy of detection and counting. Previous studies have shown that some features were used to successfully detect wheat ears from the background, including texture, color, and morphology. Narkhede et al. used color space conversion to count wheat ears by using color features in different color spaces [5]. However, the color of wheat

ears, leaves, and stalks in the wheat field are similar. Moreover, the color of various parts of the wheat plant are also changing as the wheat grows. Therefore, it was difficult to accurately identify wheat ears by using color features [6]. Some scholars have proposed counting wheat ears based on their texture and color characteristics [7,8]. However, in the heading stage, wheat ears and leaves have similar texture characteristics that affect detection accuracy. Therefore, the detection and counting of wheat ears in the natural environment still faces great challenges.

In addition, the adhesion and occlusion between wheat ears severely limits the accuracy of wheat ear identification and counting. Some scholars have successfully detected the target by the use of segmentation technology for adhesion objects, such as morphology [9], concave point matching [10], and watershed algorithm [11]. However, the morphology of wheat ears in the image is quite different. Therefore, segmentation based on morphology cannot be used to accurately count the wheat ears in the adhesion area. Moreover, the concave point matching algorithm requires that the adhered detected objects have smoother edges. However, as we know, the edges of wheat ears are not smooth, and it is difficult to obtain a smooth edge for the image of wheat ears, even if its binary image undergoes a series of erosion and expansion operations. The watershed algorithm needs to calculate local extrema. However, there are more local extremums because the texture of the wheat ear itself is clearer. Therefore, the watershed algorithm used to detect the wheat ear will lead to excessive segmentation. Although the corner points of wheat ears were used for effective segmentation to detect wheat ears [12], different wheat ears have different corner rules, which is not convenient for large-scale promotion. Therefore, how to accurately count wheat ears that are blocked by each other still needs to be solved urgently.

With the development of image processing technology, previous studies have shown that machine learning methods are used to build a classifier for wheat ear detection, thereby realizing wheat ear detection and counting [13]. Xu used the k-means algorithm to segment the wheat ears to achieve recognition [14]. Fernandez-Gallego et al. used Fourier filtering and Fourier transform to segment wheat ears and backgrounds [15]. Zhu et al. used the support vector machine method to successfully detect and count wheat ears [16]. Zhou et al. used the twin-support-vector machine segmentation method to segment and count wheat ears [17]. Although the recognition of wheat ears has been achieved based on machine learning methods, most methods still require prior knowledge to artificially set image features, which leads to insufficient robustness of features under noise interference such as uneven lighting and complex backgrounds in a field environment. Therefore, it is difficult to detect and count wheat ears in different scenarios based on traditional machine learning methods due to the lack of universality.

In the past ten years, deep learning has become a research hotspot in the field of pattern recognition. It has led to excellent achievements in many fields such as computer vision, image analysis, and multimedia applications [18]. Different from traditional pattern recognition methods, deep learning automatically learns features from big data instead of manually designed features. In fact, the characteristics of wheat ears in the field are often combined in a non-linear manner under the influence of various complex factors. The key to deep learning is to successfully separate these factors through multi-layer non-linear mapping [19]. Misra et al. integrated local patch extraction network (LPNet) and global mask refinement network (GMRNet) to achieve wheat ear segmentation and counting [20]. Xiong et al. used context-augmented local regression networks to detect and count wheat ears [21]. The research mentioned above shows that deep learning has strong robustness in detecting wheat ears. In recent years, convolutional neural network (CNN), a type of deep learning, has achieved brilliant results [22–24] in the detection and counting of wheat ears. However, the core of detection based on the CNN method is based on the region proposal method, that is, first select the sliding window or extract the proposal to train the network, and then classify it in the region proposal. The limitation of this method is that the background area is often misdetected as a specific target in object recognition. The wheat images collected in a field environment have many interferences such as high plant

density, multiple overlaps, uneven lighting, and complex backgrounds. Therefore, wheat ear detection based on deep learning still has issues worth exploring.

YOLO (you only look once) is a high-precision target detection method, which can directly predict the location and attributes of the target for the entire image based on a single convolutional network [25]. With the development of the YOLO algorithm, YOLOv4 has attracted more attention. Although YOLOv4 was successfully used to detect apple flowers [26], we still found that YOLOv4 also has insufficient bounding box positioning and it is difficult to distinguish between overlapping detected objects. The emergence of the attention mechanism can effectively solve the above problems. When processing information, the attention module only pays attention to part of the regional information that is conducive to the realization of the task, and filters out secondary information to improve the model effect, which has been used in image classification [27], image segmentation [28], and image detection [29]. Therefore, CBAM-YOLOv4 was proposed in this research, which integrated the convolutional block attention module (CBAM) [30] into the YOLOv4 convolution module to achieve the learning of target features and location features in the channel dimension and the global space dimension, respectively. The improved YOLOv4 algorithm could dynamically enhance useful features (wheat ears) and suppress background noise (wheat stalks, wheat leaves, wheat seedlings, wheat awns, and soil). As far as we know, there are few reports on the detection and counting of wheat ears using the CBAM-YOLOv4 model.

Therefore, this study focused on the feasibility of the deep learning method for high-throughput wheat ear detection and counting under field conditions, and verified that the proposed model had the ability to quickly detect wheat ears from a complex background. The purpose of this study was to (1) train a CBAM-YOLOv4 model to fine-tune the parameters of the model to achieve accurate detection of wheat ears, (2) improve the robustness of the model by using a dual-channel (channel and spatial channel) attention mechanism to eliminate background interference, and (3) realize accurate counting of wheat ears under complex backgrounds to obtain wheat high-throughput parameters, such as yield and above-ground nitrogen content, etc.

2. Materials and Methods

2.1. Data Processing

2.1.1. Data Acquisition

To make the wheat ear samples diversified, three different datasets were used in this study. Among them, the dataset of WD comes from the National Agricultural Science and Technology Innovation and Integration Demonstration Base in Anhui Province, China (31.25° N, 117.28° E). The wheat varieties are Wanmai 55 and Ningmai 15 with three nitrogen fertilizer treatments (0, 104, 150 kg/hm^2). These images were used as the original dataset and were taken with a digital camera (D5300, Nikon Corp, Tokyo, Japan). At 50–80 cm above the top of the wheat canopy, the images of the heading and maturity stages were manually taken in May 2019. A total of 190 images were selected from the collected images to construct the WD data set. All images were stored in JPG format according to the sRGB color standard, and the original resolution was 3456 × 4408 pixels. A single image contains wheat ears and leaves, and part of the wheat image is shown in Figure 1a. Among them, the wheat ear samples in these images included severe occlusion, slight occlusion, and no occlusion. The contour of a single wheat ear was incomplete due to being blocked by wheat leaves or other wheat ears. Generally, when the pixel area covered by a wheat ear in the image accounted for more than 50% of the whole wheat ear it was considered severely occluded, and less than 50% was considered lightly occluded.

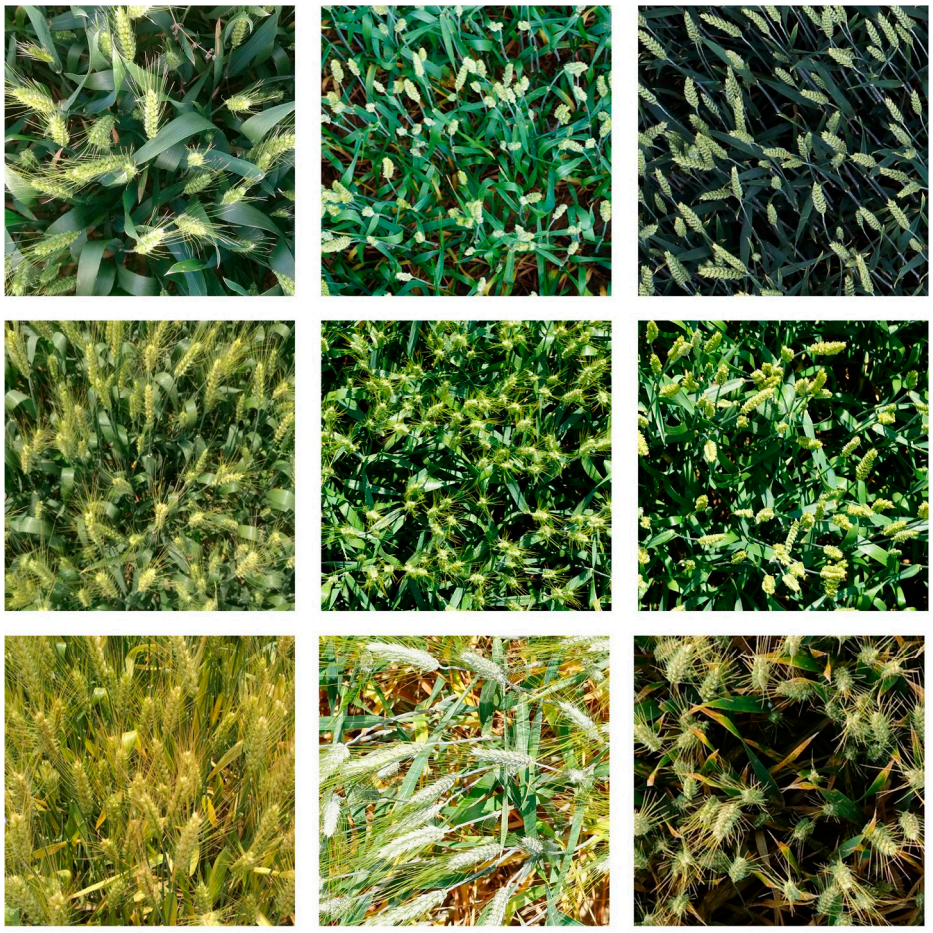


Figure 1. Samples of the wheat ears from different datasets.

The second data set was from the public data set WEDD (Wheat Ears Detection Dataset) provided by Madec et al. [31]. It contains 236 high-resolution wheat images (6000 × 4000 pixels), with a total of 30,729 wheat ears. The data was collected using a Sony ILCE-6000 digital camera, which was fixed on a boom 2.9 m from the ground for shooting. Part of the wheat image from WEDD is shown in Figure 1b.

The third data sets were from the Global Wheat Head Detection Dataset (GWHDD) [32], 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 was shown in Figure 1c.

The images in the data set were directly cropped to a size of 1024 × 1024 pixels, and included as many ear samples of wheat as possible to reduce hardware pressure and unified requirements for annotations.

The processed data sets were divided into training set, validation set and test set, as shown in Table 1. The training set was randomly sampled from the overall data set with independent and identical distribution, and the test set and the verification set were mutually exclusive, which ensured the reliability of the later evaluation standards. The validation set was used to determine the hyperparameters in the model during the training process, and the test set was used to evaluate the generalization ability of the model.

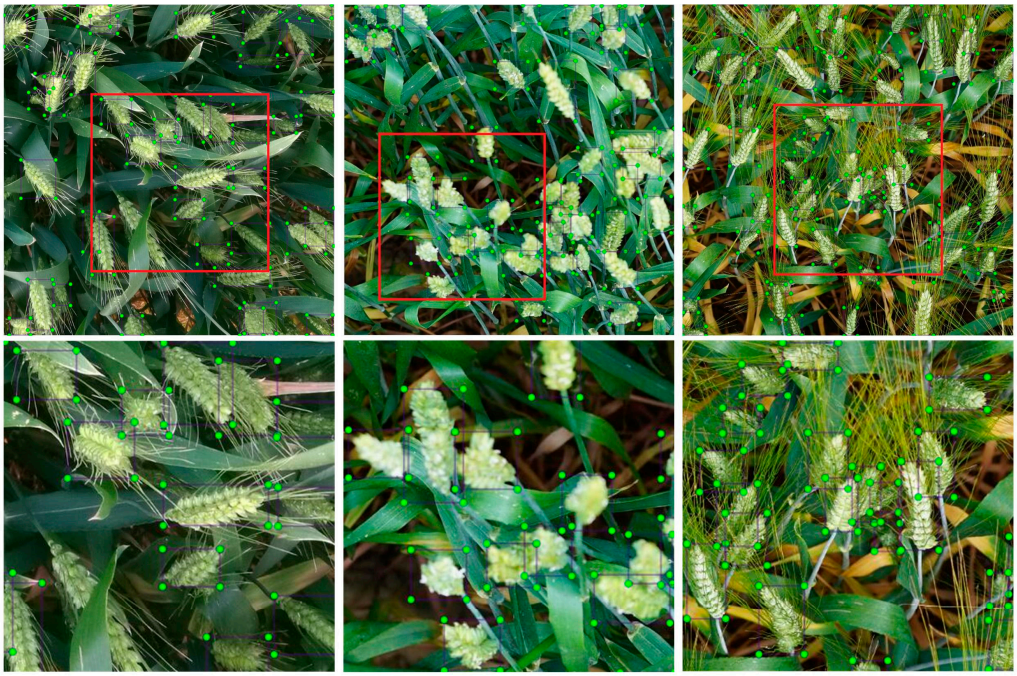
Table 1. Related datasets.

Data Sets	Image Size (Pixels)	Training Set	Validation Set	Test Set	Images
WD	3456 × 4408	133	38	19	190
WEDD	6000 × 4000	165	47	24	236
GWHDD	1024 × 1024	821	235	117	1173
Total		1119	320	160	1599

2.1.2. Data Annotation

To realize the labeling of wheat ears, labelImg (https://github.com/tzutalin/labelImg, accessed on 10 January 2021) was used to label the collected data sets. Specifically, each wheat ear in the corresponding white box of an image was annotated with a rectangular box, which was represented by the coordinates of its four vertices. After marking all the wheat ears in the corresponding frame in an image, the corresponding XML file was generated, which included information such as the size of the image, the name of the label frame, and the location of the target frame.

Figure 2 is an example of labeling of three different data sets (WD, WEDD, and GWHDD). Among them, the images in the first row of Figure 2 represent annotated samples of wheat ears. The images in the second row show a magnified portion of the annotated area in the first row of samples.



a. Sample annotation of the WD dataset

b. Sample annotation of the WEDD dataset

c. Sample annotation of the GWHDD dataset

Figure 2. Example of labeling with LabelImg.

2.1.3. 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. Methods of Wheat Ear Detection and Counting

2.2.1. YOLOv4 Model

The YOLO (You Only Look Once) network is a target detection algorithm that directly returns the position and category of the bounding box based on a convolutional neural network. The advantage of the YOLO model is that it can better distinguish the target and the background area [25]. YOLO models generally include YOLO v1 [33], YOLO v2 [34] and YOLO v3 [35]. The YOLO models mentioned above have achieved good results in many target detections. The YOLOv4 target detection algorithm is based on the YOLOV3 architecture to improve the detection accuracy of the model by optimizing data processing, backbone network, network training, activation function, and loss function [36]. Specifically, the YOLOv4 model retains the head part of yolov3, modifies the backbone network to CSPDarknet53, and uses the idea of SPP (spatial pyramid pooling) to expand the receptive field. YOLOv4 introduces a multi-scale feature extraction module to ensure strong detection performance for targets of different sizes [37]. Path Aggregation Network (PANet) mainly realizes the integration of features extracted by the backbone network. CBL (Convolution, Batch normalization, and Leaky ReLU) is the smallest component in the Yolo v4 network structure, which includes convolution, BN (Batch normalization) and Leaky ReLU functions. Compared with the YOLO3 model, the model of YOLOv4 has faster detection speed and good accuracy.

2.2.2. Channel Attention Module and Spatial Attention Module

CBAM consisted of a channel attention module and spatial attention module, as shown in Figure 3. For input feature F, the global information of each feature channel was obtained through global average pooling and maximum pooling operations, and then the feature channel attention vector was obtained through two fully connected layers, FC1 and FC2, which was used to weight the input feature F channel by channel to obtain the feature F′.

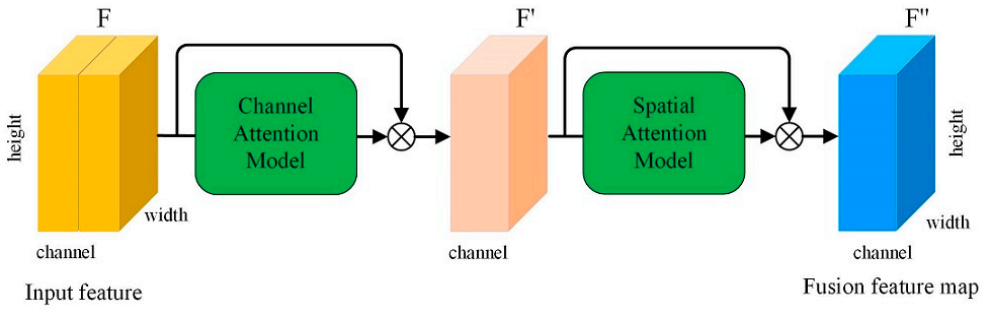


Figure 3. Structure illustration of the channel and spatial attention module.

For the spatial attention module, the feature is subjected to the maximum pooling operation to obtain the feature map, which is 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″ [30].

The information for detecting wheat ear characteristics was usually concealed by leaves or other wheat ears. Therefore, the channel attention module could enhance the feature expression of the occluded target, and the spatial attention module could highlight areas in the feature map that are related to the current task.

2.2.3. Model of Wheat Ear Detection and Counting Based on CBAM-YOLOv4

As shown in Figure 4, the process of wheat ear detection and counting model mainly included two modules; one was the batch training module, the other was the detection and counting module. Among them, the batch training module included original data samples,

data expansion, data labeling, divided data sets, and CBAM-YOLOv4 model construction; the detection and counting module includes wheat ear detection and counting.

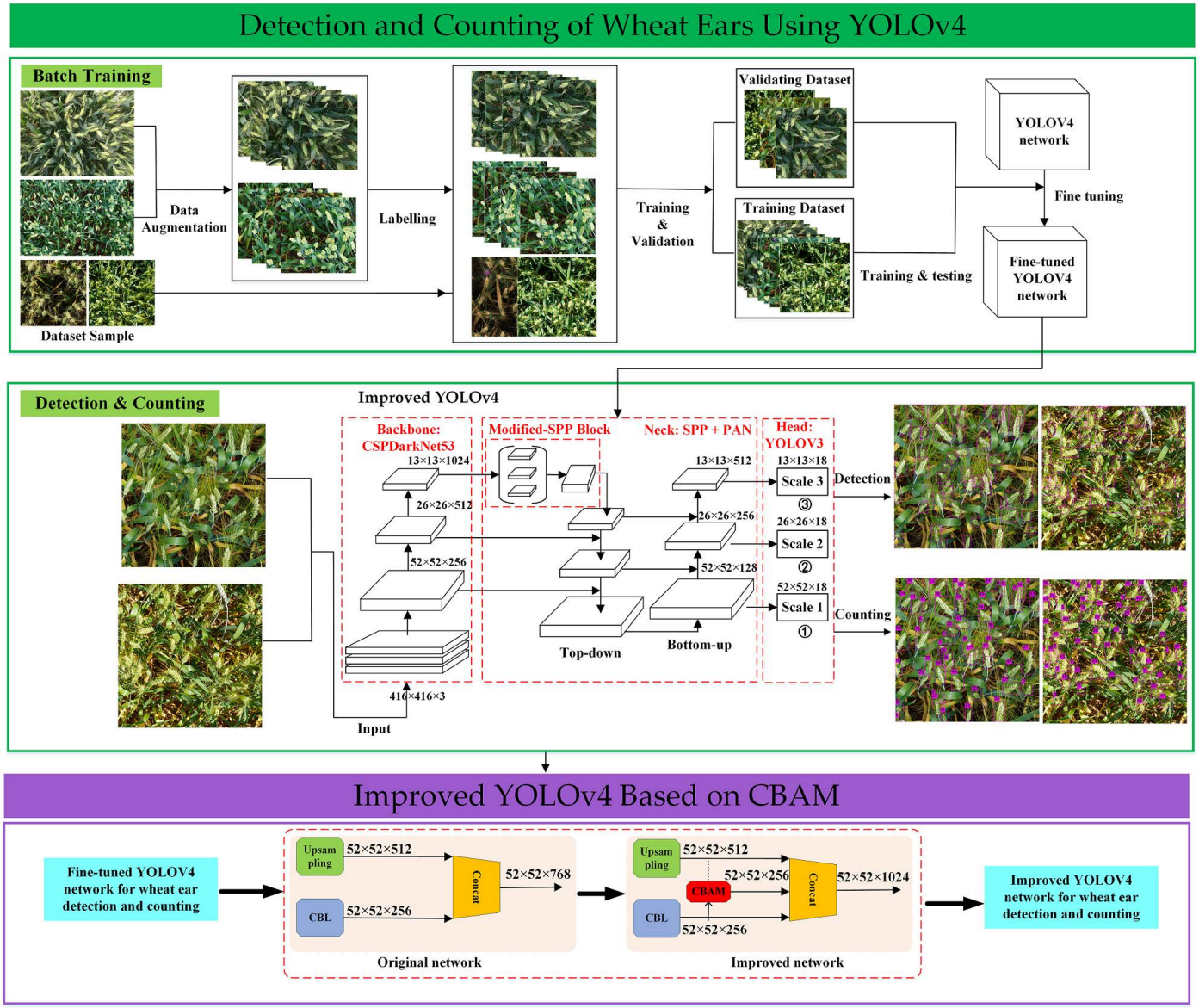


Figure 4. Technical flow chart of wheat ear detection and counting based on improved YOLO v4.

The detection and counting module were as follows: firstly, the training set and the validation set were constructed using the manually labeled wheat ear data set. Secondly, the training sets of wheat images were used to fine-tune the model based on the transfer learning method. Once more, the CBAM-YOLOv4 model was further adjusted, optimized, and verified using the verification set. Finally, the model was tested using the test set, and the detection and counting results of wheat ears were generated. Among them, the wheat ear recognition results were presented in the form of a bounding box. Counting wheat ears was based on wheat ear identification, and the results are displayed with a bounding box and statistical serial numbers.

According to the structure of the YOLOv4 detection model, we embed the CBAM in the neck area of YOLOv4, and the result is shown in Figure 5. Among them, the image size of the wheat input to the model is 416 × 416 × 3, and the output feature maps A1, B1, and C1 of CSPDarknet53 through the SPP network and the CBAM of F1, F2, and F3 to generate feature maps A2, B2, and C2, containing feature attention mechanisms, respectively, with

sizes of 52 × 52 × 256, 26 × 26 × 512, and 13 × 13 × 1024. Then, after outputting feature maps A2, B2, and C2 through the CBL × 5 module of the PAN network, feature maps A3, B3, and C3 will be generated respectively, with sizes of 52 × 52 × 128, 26 × 26 × 256, 13 × 13 × 512. Figure 5 shows the modification area of CBAM-YOLOv4.

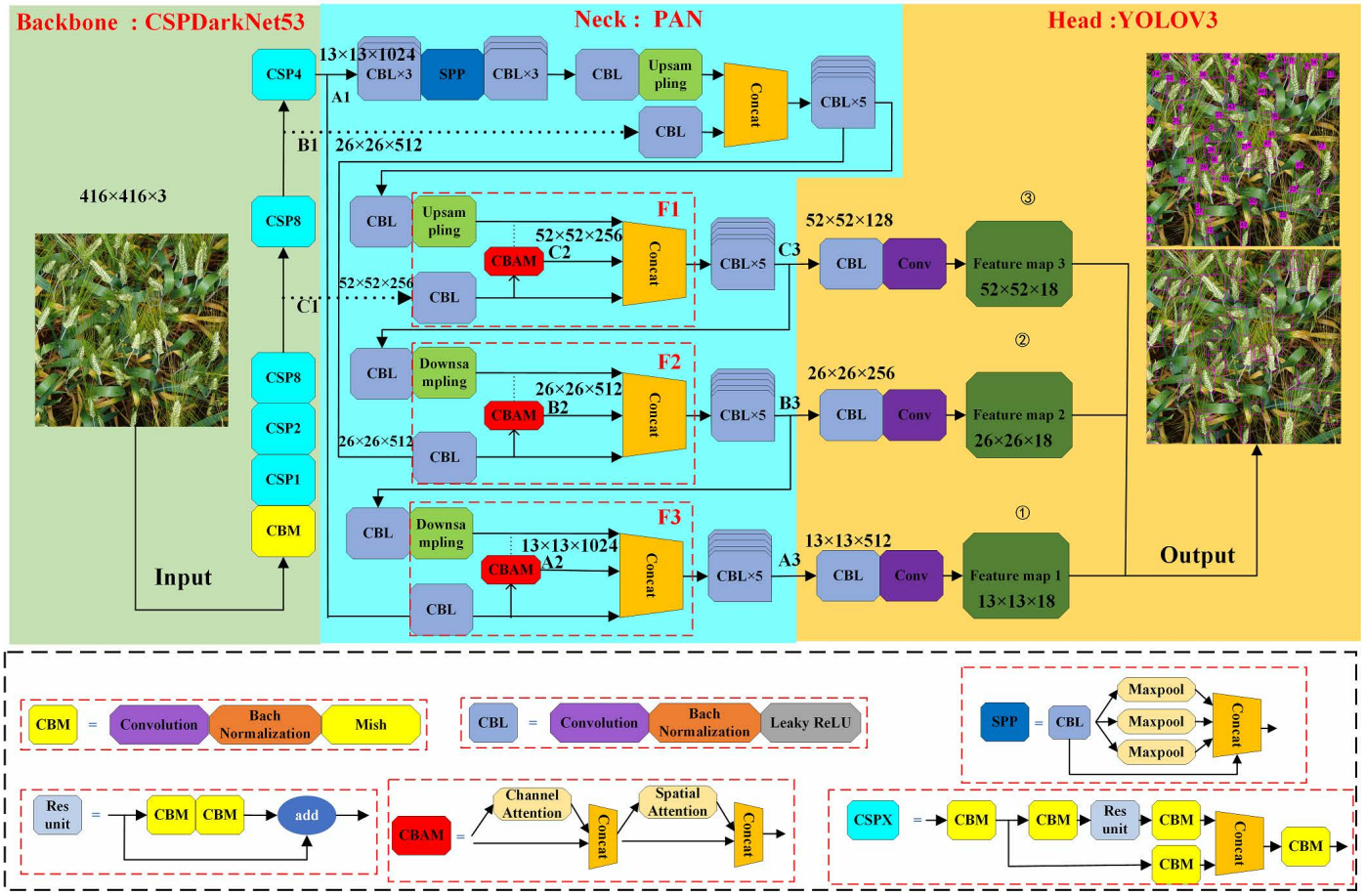


Figure 5. YOLOV4 model with spatial and channel attention module.

2.2.4. Method of Wheat Ear Detection and Counting

The detection and counting results of wheat ears are presented in the form of a detection box, and the counting includes counting the target value of the wheat ears. The counting method is: traverse all the detection frames and set a threshold N. When S > N, it is considered to belong to the wheat ear target, and then the detection frame participates in the statistics. When S ≤ N, it is considered not to belong to the wheat ear target. This detection frame does not participate in statistics. The specific value is displayed in the detection box participating in the statistics (starting with the number 1, and accumulating in sequence). The formula is as shown:

$$S_x = \begin{cases} 1, & S_x > N \\ 0, & S_x \leq N \end{cases} \quad (1)$$

where: S_x is the score of the x-th detection frame, and N is the set threshold, the threshold N = 0.5 in this study.

In addition, we asked eight scholars in the field to manually count the wheat ears from the image of the test set ten times, which was averaged as the ground truth value of the wheat ears, and CSRNet [38] was used to visualize the distribution of the number of wheat ears.

2.2.5. Evaluation of the Model Performance

To test the effectiveness of the CBAM-YOLOv4 model and to verify the transfer performance of the attention information on the model, intersection over union (IOU) is 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 is 0.5 in this study.

In addition, the precision, recall, F1-score, and mean average precision (mAP) were used as evaluation indicators to evaluate the trained model:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{3}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{mAP} = \frac{1}{C} \sum_{k=i}^{N} P(k) \Delta R(k) \tag{5}$$

Among them, true positives (TP) means that both the detection result and the true value are wheat ears, that is, the number of wheat ears 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) means 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 is used to evaluate the performance of the method by balancing the weights of precision and recall. C is the number of categories, N represents the number of all pictures in the test set, $P(k)$ represents the Precision when k pictures can be recognized, and $\Delta R(k)$ represents the change of the recall value when the number of recognized pictures changes from $k-1$ to k.

Coefficient of determination (R^2), root mean square error (RMSE) and Bias are used as evaluation indicators to measure the counting performance of the model, which are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (m_i - c_i)^2}{\sum_{i=1}^{n} \left(m_i - \overline{m} \right)^2} \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (m_i - c_i)^2} \tag{7}$$

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (m_i - c_i) \tag{8}$$

where n represents the number of wheat ear pictures, m_i, 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 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,

max batches = 10,000, momentum = 0.9, decay = 0.0005, batch size = 16. Mini-batch gradient descent (MBGD) was used to optimize the training model. The hardware parameters of the experiment were Intel Core i7-8700 processor and NVIDIA GeForce GTX 2080 GPU, which were implemented using Darknet deep learning framework and Python programming. CUDA version 10.0 parallel computing framework and CUDNN version 7.5 deep neural network acceleration library were used in this study.

Figure 6 shows the model verification accuracy and training loss values obtained in each iteration during the training process. It can be seen from Figure 6 that the training accuracy of the model gradually increases as the number of iterations increases, and the training loss value of the model gradually decreases as the number of iterations increases. In the initial stage of model training, the model learning efficiency was high, and the training loss curve converges more quickly. As the number of iterations increases, the slope of the training loss curve gradually decreases. Finally, when the number of training iterations reaches about 8500, the fluctuation trend of the loss value gradually stabilizes, and the corresponding accuracy no longer changes. Among them, the maximum value of mAP was 88.76%, indicating that the CBAM-YOLOv4 model did not have problems such as over-fitting or under-fitting, and gradient disappearance. This model was the model used to detect and count wheat ears.

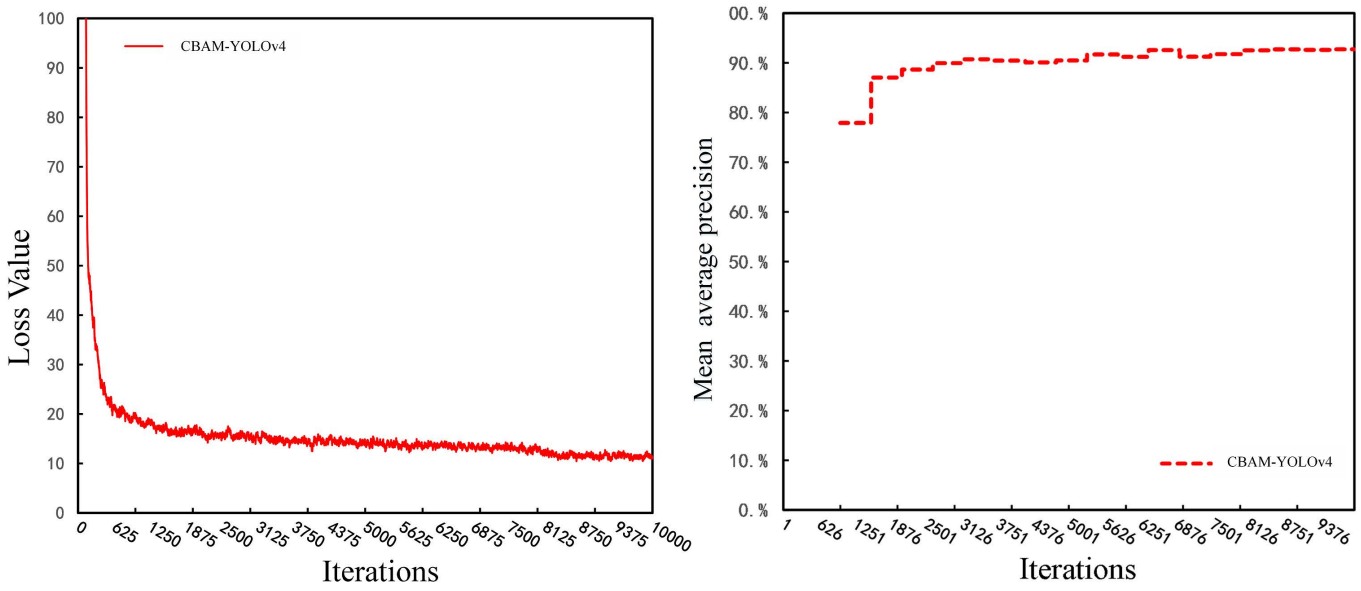


Figure 6. Loss value and mean average precision changes with iterations.

3.2. Results of Detecting Wheat Ears in Different Data Sets

The trained CBAM-YOLOv4 model was tested using test sets from three data sets. There were 771 images in the test set, including 30 images in the WD test set, 48 images in the WEDD test set, and 673 images in the GWHDD test set. The resolution of the smallest wheat ear was no less than 15 × 15 pixels, and it was ignored if the size was too small.

It could be seen from Table 2 that CBAM-YOLOv4 was effective in detecting wheat ears in different data sets, even though the shape, color, and texture of wheat ears in the image were different. For WD, WEDD and GWHDD, the F1-score was 88.89%, 93.02%, and 89.25%, the mAP was 94%, 96.04%, and 93.11%, the Precision was 86.06%, 89.73%, and 87.55%, and recall was 91.91%, 96.55%, and 91.01%, respectively.

Part of the test results were shown in the Figure 7. Judging from the detection results, the CBAM-YOLOv4 model could detect wheat ears in different data sets. It can be seen from Figure 7 that the WEDD detection effect was best in the three data sets, and the detection accuracy of the WD and GWHDD test sets were a little lower than that of WEDD.

Table 2. Evaluation indicators of wheat ear detection of three data sets.

Different Data	F1-Score/%	Precision	Recall	mAP
WD	88.89%	86.06%	91.91%	94.00%
WEDD	93.02%	89.73%	96.55%	96.40%
GWHDD	89.25%	87.55%	91.01%	93.11%

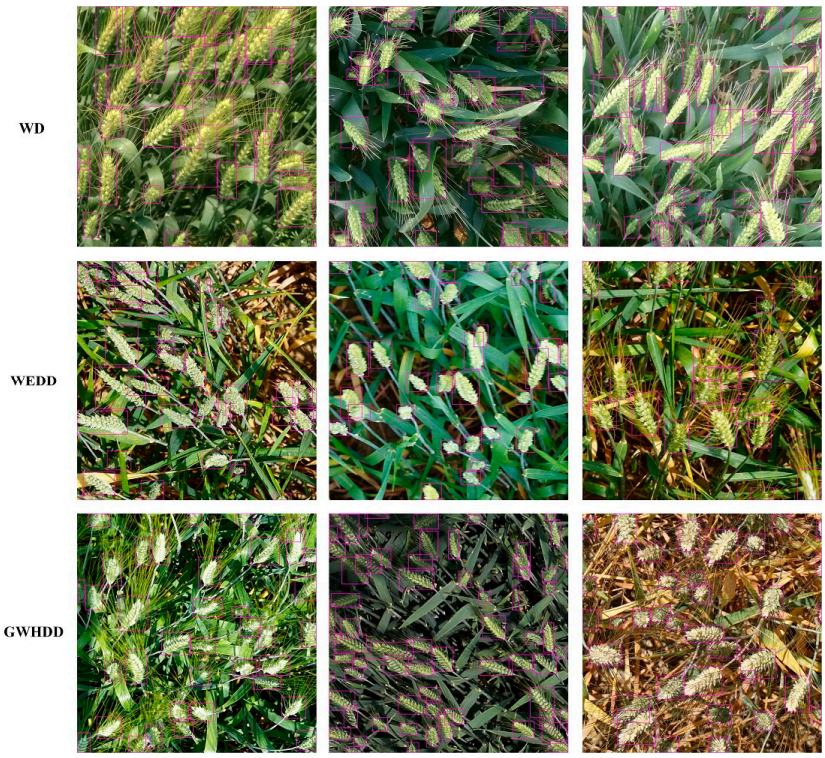


Figure 7. Examples of wheat ear detection results of different data sets.

3.3. Results of Counting Wheat Ears in Different Data Sets

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 WD, WEDD, and GWHDD, and 617, 535, and 742 wheat ears were manually counted on the images. The CBAM-YOLOv4 model was used to detect and count the number of wheat ears. The results are shown in Figure 8. The R^2 of the model was 0.8968 for WD, 0.9550 for WEDD, and 0.9884 for GWHDD. The RMSE corresponding to the three data sets are 1.604, 1.070, and 2.097. In addition, for WD, WEDD, and GWHDD, the results of CBAM-YOLOv4 method count deviation values are 2.6, −1.95, and −3.3, respectively. Moreover, compared with the actual number of wheat ears, the detection result of CBAM-YOLOv4 had a certain deviation. The Bias were 2.6, −1.95, and −3.3 for WD, WEDD and GWHDD, respectively.

Figure 9 shows part of the counting results from the three data sets, which were counted based on the method we proposed, and the results were compared with the ground truth. It could be seen that it was easy to see that the method we proposed had better robustness, and even if the distribution of wheat ears were relatively concentrated, better counting results could be obtained. At the same time, we also found that the improved YOLOv4 method could detect wheat ears with complex background, but some of the detection results were higher than the ground truth. The possible reason was that the sample label was insufficient, which led to some wheat leaves being misjudged.

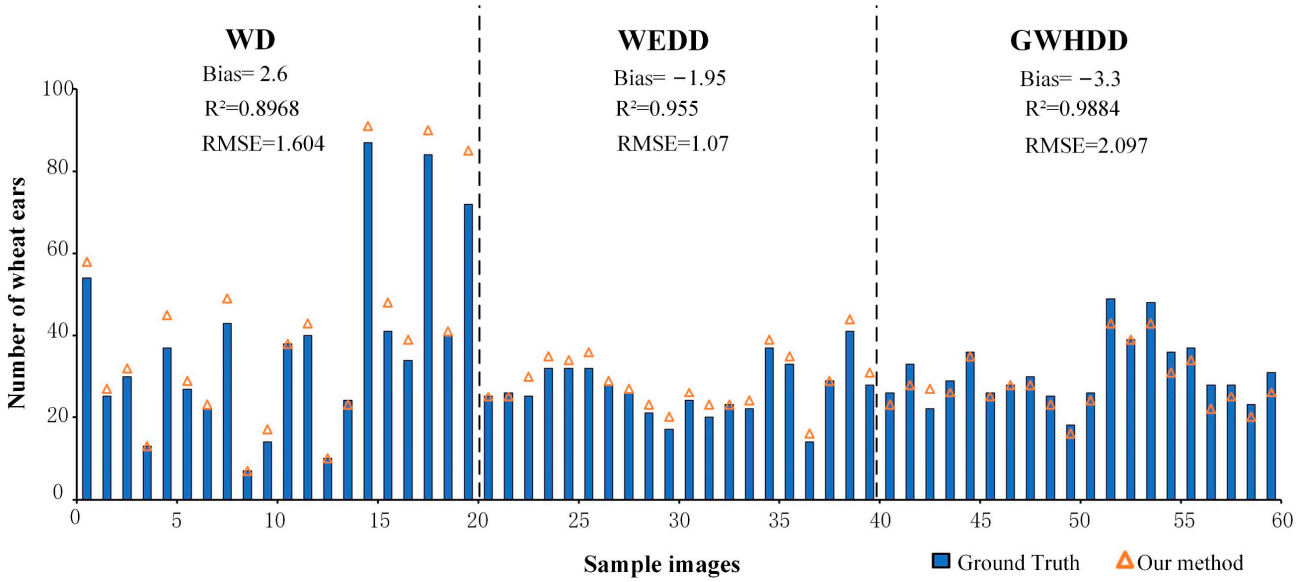


Figure 8. The comparison between the true value and the estimated value of a single image of different data sets.

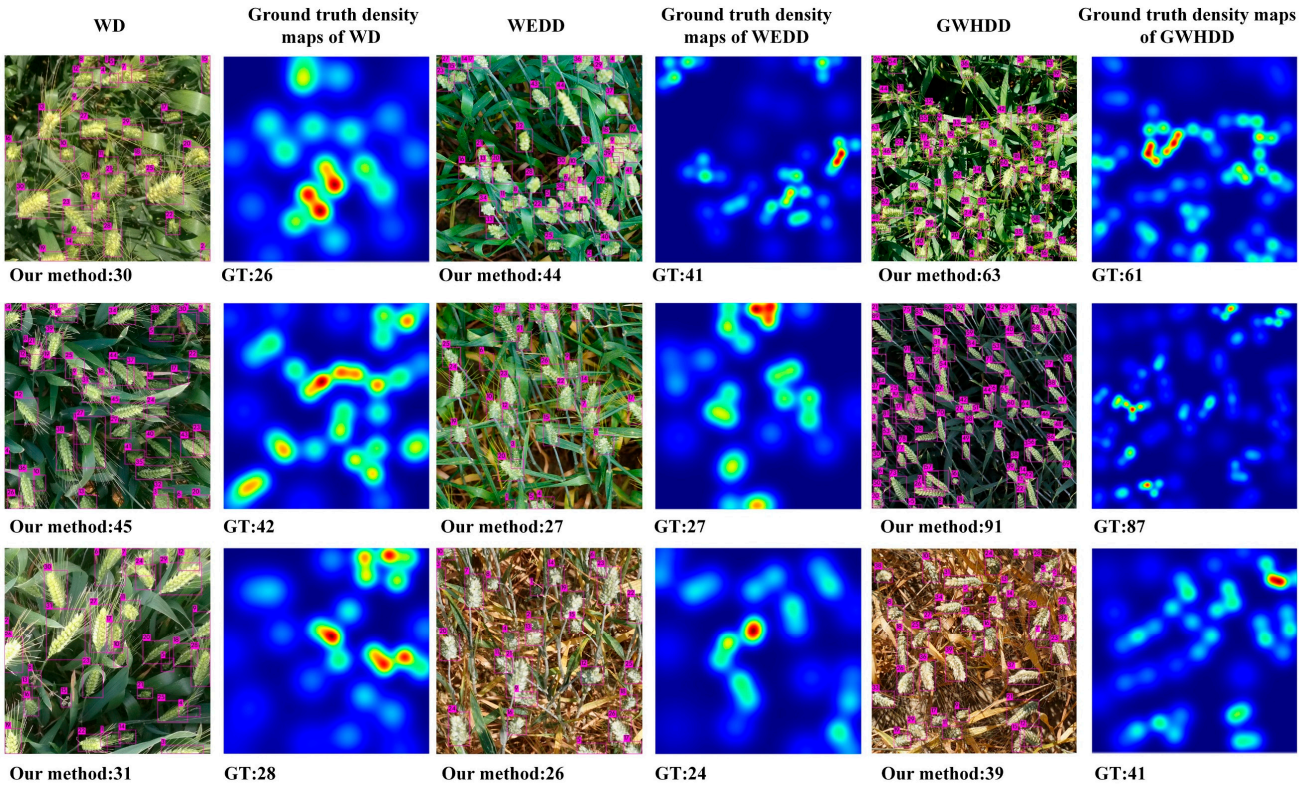


Figure 9. The comparison between the ground truth and the counting with our method. GT: ground truth.

4. Discussion

4.1. Comparison of the Effect of Wheat Ear Detection and Counting under 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 cover the samples and the wheat ear samples overlap 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 from WD, and 20 images with slightly occluded wheat ears as data set B from WD. The degree of occlusion

was used as a control variable, and the CBAM-YOLOv4 model was used to detect data sets A, B, and A + B, respectively. The detection results are shown in Table 3 and Figure 10. For the detection of lightly obscured wheat ears (data sets A), the F1-score of the model can reach 0.9253, and the mAP can reach 95.38%. In a heavily occluded environment with dense targets (data sets B), the model can also achieve an F1 value of 0.9019 and the mAP value of 96.25%. The two data sets were mixed into one data set (A + B), and the F1-score and the mAP of the model reached 0.9124 and 93.92%. It showed that the CBAM-YOLOv4 model could effectively detect wheat ears in the natural field environment.

Table 3. Comparison of detection results of wheat ears with different degrees of occlusion.

Test Set	Precision	Recall	F1-Score	mAP (%)
A	0.9201	0.9305	0.9253	95.38
B	0.9209	0.8823	0.9019	92.65
A + B	0.9205	0.9045	0.9124	93.92

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

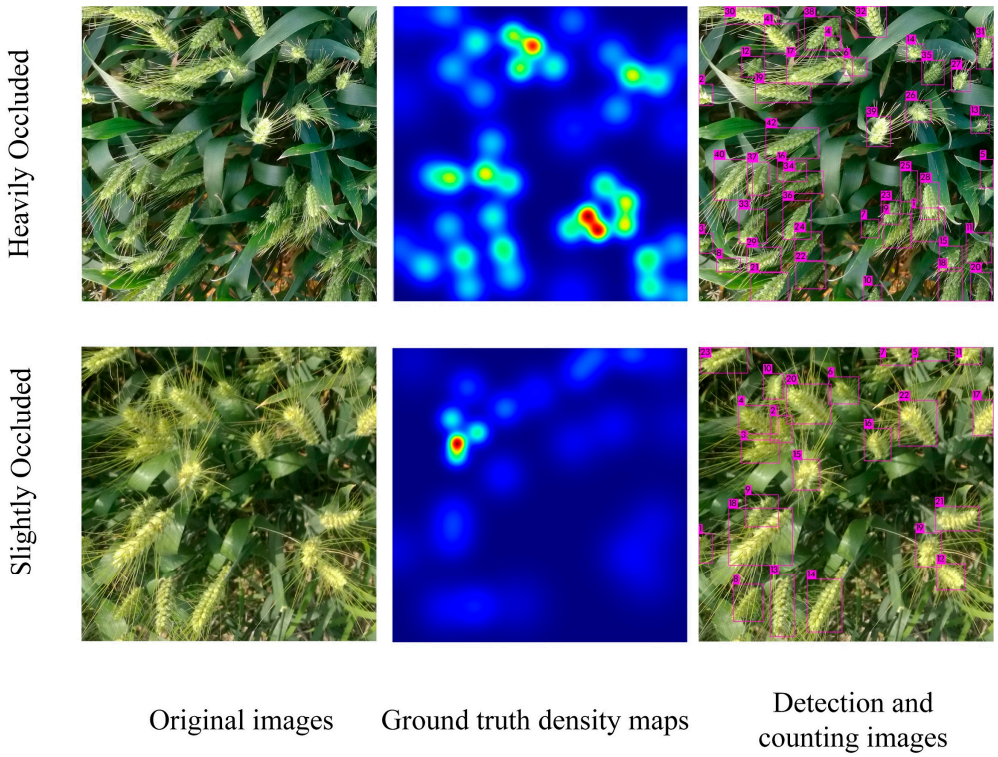


Figure 10. Comparison of detection results of wheat ears with different degrees of occlusion.

As shown in the density maps in Figure 10, the colors in the maps reflect the size of the density value. The darker the color, the greater the density value. It could be seen from Figure 10 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 planting of wheat will result in denser wheat ears and severe occlusion. On the other hand, the camera shooting angle will also increase the sample occluded by wheat ears [30]. In fact, in fields with dense wheat ears, even experts must count the number of wheat ears multiple times to obtain reliable measurement results.

4.2. Comparison of Detection Effects Based on Different YOLO Methods

To evaluate the performance of the CBAM-YOLOv4 model proposed in this study, the results of the detection of wheat ears by typical convolutional neural networks YOLOV3,

YOLOv4, and CBAM-YOLO4 were compared using the test set. Some examples of the results of the GWHDD test set are shown in Figure 11. The rectangular box in Figure 11 is the result of detecting wheat ears, the red circle represents the result of false detection of wheat ears, and the blue circle represents the result of missing wheat ears. When comparing the detection of wheat ears in the same image, it can be seen from Figure 11 that we found that there were missed ear detections and false detections of wheat ears in the results based on YOLOV3 and YOLOv4 model detection, but the CBAM-YOLOv4 model can accurately detect wheat ears, which showed that the model had good robustness, and that the attention information was beneficial to the detection of wheat ear targets.

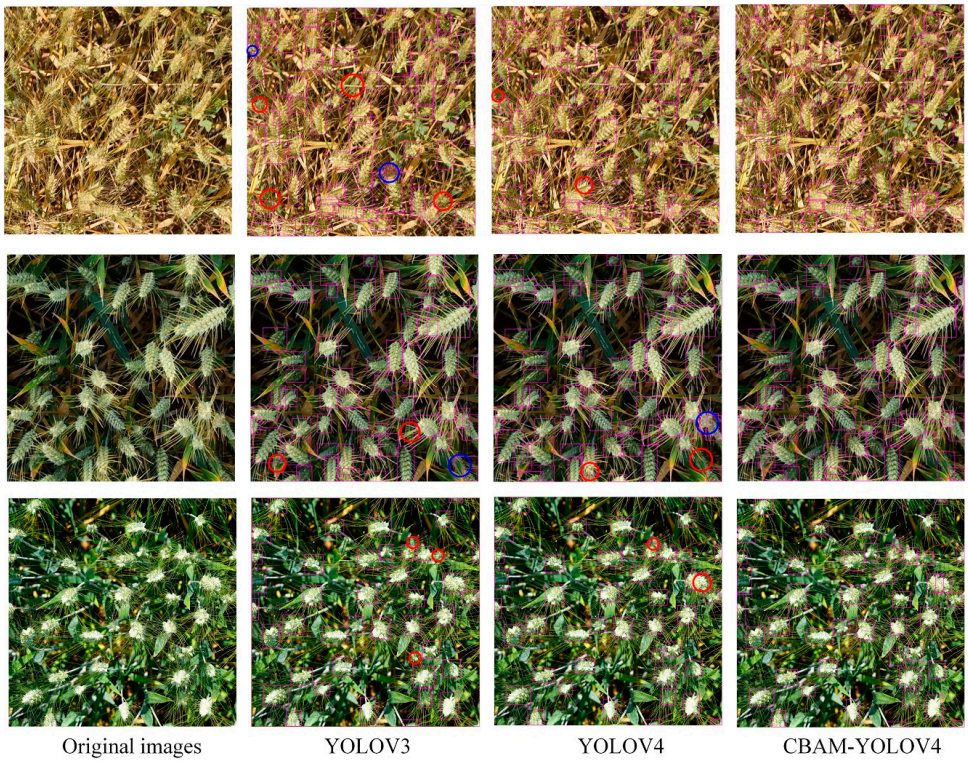


Figure 11. Examples of the results of different algorithms detecting wheat ears with the test sets.

In addition, it could be seen from Table 4 that the precision, recall, F1-score, and mAP of the CBAM-YOLO4 model were the highest. Among them, the mAP of CBAM-YOLO4 was 93.11%, which was 3.89% and 1.98% higher than that of the YOLOV3 and YOLOv4 models, respectively. The results showed that the detection effect of the CBAM-YOLO4 model was better than that of the YOLOV3 and YOLOv4 models.

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

Model	Precision	Recall	F1-Score	mAP (%)
YOLOV3	0.8204	0.8932	0.8553	89.48
YOLOv4	0.8577	0.9013	0.879	91.26
CBAM-YOLOv4	0.8755	0.9101	0.8925	93.11

Figure 12 showed the precision–recall curves of three YOLO models on wheat ears. It can be seen from Figure 12 that when the recall of the three models were less than 0.1, the precision remained around 1.0, and the difference was not significant. However, with the increase in the recall value, the advantages of the CBAM-YOLOv4 model gradually become obvious, and 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 the spatial attention module.

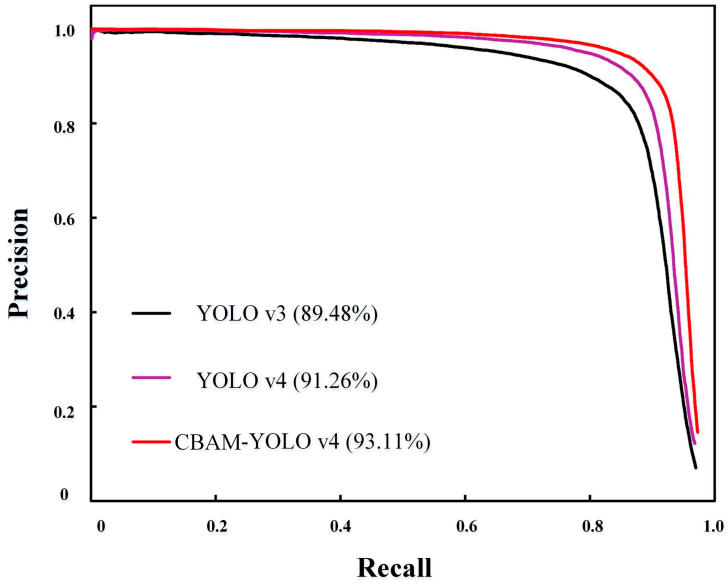


Figure 12. Precision–recall curves of the detection results of different algorithms with test set.

4.3. Estimation of Yield and Aboveground Nitrogen Content Based on of Wheat Ears

The number of wheat ears detected, not only allows us to quickly predict wheat yield, but also to easily obtain other high-throughput parameters of wheat such as above-ground nitrogen content. Aboveground nitrogen content (ANC, kg·ha^{-1}) of wheat is determined by Equation (9) [39]:

$$ANC = \frac{SPNC \times m \times \#Ears}{k} \quad (9)$$

where SPNC is the sample plant nitrogen concentration (g·100 g^{-1}), m is the dry mass (kg·ha^{-1}), k is the number of samples and *#Ears* is the number of wheat ears. In previous studies, when we calculated ANC, the number of wheat ears was manually collected. Now, we can use images to count wheat ears in a certain area in the field.

Moreover, the method in this study was used to quickly detect and count the number of wheat ears. Based on the number of wheat ears, other high-throughput parameters of wheat grains, such as starch content and nitrogen content, can quickly be obtained by calculation, the details are shown in Figure 13.

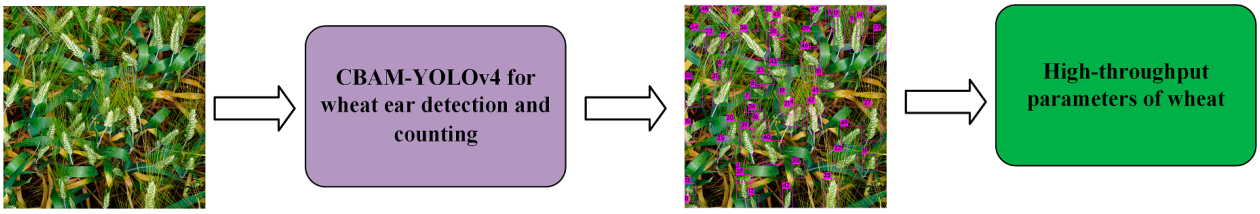


Figure 13. Compare the results of wheat ear counts with different density distributions.

5. Conclusions

We combined the convolutional neural network and attention mechanism technology to propose a CBAM-YOLOv4 wheat ear detection and counting method. The model was trained, validated, and tested using public data sets (WEDD and GWHDD) and data sets (WD) obtained by ourselves. The key contribution was that the CBAM-YOLOv4 model had good robustness. By being integrated into the dual-channel attention mechanism, the YOLOv4 model paid more attention to important wheat ear features in the image, and suppressed unimportant features such as wheat leaves and wheat awns. Thus, the CBAM-YOLOv4 model improves the accuracy of detecting and counting wheat ears. Moreover,

the model was verified and tested using wheat ear images from different countries and regions. The mAP of the method proposed in this study exceeded 91%, meaning that it can effectively detect and count wheat ears.

Author Contributions: Methodology, B.Y. and Y.G.; software, Y.G.; data curation, Z.G. and Y.Z.; writing—review and editing, B.Y.; visualization, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Anhui Province (1808085MF195), the Opening Project of Key Laboratory of Power Electronics and Motion Control of Anhui Higher Education Institutions (PEMC2001), the Open Fund of State Key Laboratory of Tea Plant Biology and Utilization (SKLTOF20200116), and the Open Fund of the Key Laboratory of Technology Integration and Application in Agricultural Internet of Things, the Ministry of Agriculture (2016KL02).

Data Availability Statement: Restrictions apply to the availability of these data. The data were obtained from the Smart Agriculture Research Institute of Anhui Agricultural University and are available from the authors with the permission of the Smart Agriculture Research Institute.

Acknowledgments: We would like to thank Lin Qi and Mengxuan Wang for their help with field data collection. 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. Fang, Y.; Qiu, X.L.; Guo, T.; Wang, Y.Q.; Cheng, T.; Zhu, Y.; Chen, Q.; Cao, W.X.; Yao, X.; Niu, Q.S.; et al. An automatic method for counting wheat tiller number in the field with terrestrial lidar. *Plant Methods* **2020**, *16*, 132. [CrossRef]
3. Fernandez-Gallego, J.A.; Buchaillot, M.L.; Aparicio Gutiérrez, N.; Nieto-Taladriz, M.T.; Araus, J.L.; Kefauver, S.C. Automatic Wheat Ear Counting Using Thermal Imagery. *Remote Sens.* **2019**, *11*, 751. [CrossRef]
4. Qiu, R.; Wei, S.; Zhang, M.; Li, H.; Li, M. Sensors for measuring plant phenotyping: A review. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 1–17. [CrossRef]
5. Narkhede, P.R.; Gokhale, A.V. Color image segmentation using edge detection and seeded region growing approach for CIELab and HSV color spaces. In Proceedings of the 2015 International Conference on Industrial Instrumentation and Control (ICIC), Pune, India, 28–30 May 2015; pp. 1214–1218.
6. Germain, C.; Rousseaud, R.; Grenier, G. Non destructive counting of wheatear with picture analysis. In Proceedings of the Fifth International Conference on Image Processing and its Applications, Edinburgh, UK, 4–6 July 1995; pp. 435–439.
7. Cointault, F.; Guerin, D.; Guillemin, J.P.; Chopinet, B. In-field Triticum aestivum ear counting using colour-texture image analysis. *N. Z. J. Crop Hortic.* **2008**, *36*, 117–130. [CrossRef]
8. Li, Q.Y.; Cai, J.H.; Berger, B.; Okamoto, M.; Miklavcic, S.J. Detecting spikes of wheat plants using neural networks with Laws texture energy. *Plant Methods* **2017**, *13*, 83.
9. 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]
10. Li, Y.; Du, S.; Yao, M.; Yi, Y.; Yang, J.; Ding, Q.; He, R. Method for wheatear counting and yield predicting based on image of wheatear population in field. *Nongye Gongcheng Xuebao Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 185–194.
11. 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]
12. Du, Y.; Cai, Y.; Tan, C.W.; Li, Z.; Yang, G.; Feng, H.; Dong, H. Field wheat ears counting based on superpixel segmentation method. *Sci. Agric. Sin.* **2019**, *52*, 21–33.
13. 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.
14. 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]
15. Fernandez-Gallego, J.A.; Kefauver, S.C.; Gutiérrez, N.; Nieto-Taladriz, M.; Araus, J. Wheat ear counting in-field conditions: High throughput and low-cost approach using RGB images. *Plant Methods* **2018**, *14*, 22. [CrossRef] [PubMed]
16. Zhu, Y.J.; Cao, Z.G.; Lu, H.; Li, Y.N.; Xiao, Y. In-field automatic observation of wheat heading stage using computer vision. *Biosyst. Eng.* **2016**, *143*, 28–41. [CrossRef]
17. Zhou, C.Q.; Liang, D.; Yang, X.D.; Yang, H.; Yue, J.B.; Yang, G.J. Wheat ears counting in field conditions based on multi-feature optimization and TWSVM. *Front. Plant Sci.* **2018**, *9*, 1024. [CrossRef] [PubMed]

18. Jermsittiparsert, K.; Abdurrahman, A.; Siriattakul, P.; Sundeeva, L.A.; Maseleno, A. Pattern recognition and features selection for speech emotion recognition model using deep learning. *Int. J. Speech Technol.* **2020**, *23*, 1–8. [CrossRef]
19. Pearline, S.A.; Kumar, V.S.; Harini, S. A study on plant recognition using conventional image processing and deep learning approaches. *J. Intell. Fuzzy Syst.* **2019**, *36*, 1997–2004. [CrossRef]
20. Misra, T.; Arora, A.; Marwaha, S.; Chinnusamy, V.; Rao, A.R.; Jain, R.; Sahoo, R.N.; Ray, M.; Kumar, S.; Raju, D. SpikeSegNet-a deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods* **2020**, *16*, 40. [CrossRef] [PubMed]
21. 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]
22. 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*, 100. [CrossRef] [PubMed]
23. 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]
24. 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]
25. Liu, G.X.; Nouaze, J.C.; Touko Mbouembe, P.L.; Kim, J.H. YOLO-Tomato: A Robust Algorithm for Tomato Detection Based on YOLOv3. *Sensors* **2020**, *20*, 2145. [CrossRef]
26. Wu, D.; Lv, S.; Jiang, M.; Song, H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.* **2020**, *178*, 105742. [CrossRef]
27. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
28. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 325–341.
29. Ba, R.; Chen, C.; Yuan, J.; Song, W.; Lo, S. SmokeNet: Satellite Smoke Scene Detection Using Convolutional Neural Network with Spatial and Channel-Wise Attention. *Remote Sens.* **2019**, *11*, 1702. [CrossRef]
30. 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.
31. Madec, S.; Jin, X.; Lu, H.; De, S.; Liu, S.; Duyme, F.; Heritier, E.; Baret, F. Ear density estimation from high resolution RGB imagery using deep learning technique. *Agric. For. Meteorol.* **2019**, *264*, 225–234. [CrossRef]
32. 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]
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA; 2016; pp. 779–788.
34. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
35. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
36. Silva, L.A.; Blas, H.; García, D.P.; Mendes, A.S.; Villarrubia, G. An architectural multi-agent system for a pavement monitoring system with pothole recognition in UAV images. *Sensors* **2020**, *20*, 6205. [CrossRef] [PubMed]
37. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLO v4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
38. Li, Y.; Zhang, X.; Chen, D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
39. Yang, B.; Wang, M.; Sha, Z.; Wang, B.; Chen, J.; Yao, X.; Cheng, T.; Cao, W.; Zhu, Y. Evaluation of aboveground nitrogen content of winter wheat using digital imagery of unmanned aerial vehicles. *Sensors* **2019**, *19*, 4416. [CrossRef] [PubMed]