frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpls.2022.821717

Wheat Spike Detection and Counting in the Field Based on SpikeRetinaNet

Changji Wen[1,2†], Jianshuang Wu[3], Hongrui Chen[3], Hengqiang Su[1,2], Xiao Chen[1,2], Zhuoshi Li[1,4] and Ce Yang[3†]*

[1] College of Information and Technology, Jilin Agricultural University, Changchun, China, [2] Institute for the Smart Agriculture, Jilin Agricultural University, Changchun, China, [3] College of Food, Agricultural and Natural Resource Sciences, University of Minnesota, Paul, MN, United States, [4] Key Laboratory of Bionic Engineering, Ministry of Education, Jilin University, Changchun, China

The number of wheat spikes per unit area is one of the most important agronomic traits associated with wheat yield. However, quick and accurate detection for the counting of wheat spikes faces persistent challenges due to the complexity of wheat field conditions. This work has trained a RetinaNet (SpikeRetinaNet) based on several optimizations to detect and count wheat spikes efficiently. This RetinaNet consists of several improvements. First, a weighted bidirectional feature pyramid network (BiFPN) was introduced into the feature pyramid network (FPN) of RetinaNet, which could fuse multiscale features to recognize wheat spikes in different varieties and complicated environments. Then, to detect objects more efficiently, focal loss and attention modules were added. Finally, soft non-maximum suppression (Soft-NMS) was used to solve the occlusion problem. Based on these improvements, the new network detector was created and tested on the Global Wheat Head Detection (GWHD) dataset supplemented with wheat-wheatgrass spike detection (WSD) images. The WSD images were supplemented with new varieties of wheat, which makes the mixed dataset richer in species. The method of this study achieved 0.9262 for mAP50, which improved by 5.59, 49.06, 2.79, 1.35, and 7.26% compared to the state-of-the-art RetinaNet, single-shot multiBox detector (SSD), You Only Look Once version3 (Yolov3), You Only Look Once version4 (Yolov4), and faster region-based convolutional neural network (Faster-RCNN), respectively. In addition, the counting accuracy reached 0.9288, which was improved from other methods as well. Our implementation code and partial validation data are available at https://github.com/wujians122/The-Wheat-Spikes-Detecting-and-Counting.

Keywords: wheat spikes, detection and counting, deep learning, attentional mechanism, wheat yield

OPEN ACCESS

Edited by:
Wanneng Yang,
Huazhong Agricultural University, China

Reviewed by:
Xiaohu Zhang,
Nanjing Agricultural University, China
Alka Arora,
Indian Agricultural Statistics Research Institute, Indian Council of Agricultural Research, India
Lingfeng Duan,
Huazhong Agricultural University, China

***Correspondence:**
Changji Wen
Changjiw@jlau.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *24 November 2021*
Accepted: *17 January 2022*
Published: *03 March 2022*

Citation:
Wen C, Wu J, Chen H, Su H, Chen X, Li Z and Yang C (2022) Wheat Spike Detection and Counting in the Field Based on SpikeRetinaNet. Front. Plant Sci. 13:821717. doi: 10.3389/fpls.2022.821717

INTRODUCTION

As one of the three major cereal crops, wheat provides food for approximately one-third of the world's population. Global wheat consumption has increased due to rising per capita income and urbanization. On the other hand, wheat crops are increasingly being hampered by phenological changes, shrinking germplasm areas, and other stresses. Therefore, wheat genetic improvement is critical to address future food security. At present, most wheat cultivation and breeding researchers rely on costly manual counting. This time-consuming process is driving the need for new tools. In addition, subjectivity and fatigue will lead to mistakes in counting wheat spikes (Jin et al., 2017). When assessing crop genetic improvement, although genotyping is easier and more accurate than

before, efficient phenotyping algorithms and techniques still limit the establishment of phenotype-genotype relationships (Eversole et al., 2014). Therefore, the construction of efficient phenotypic algorithms and technologies are particularly urgent and necessary for improving genetic efficiency. Furthermore, wheat yield is one of the important indexes of quality breeding. So, the detection and counting of spikes efficiently are one of the main research directions of phenotypic technology based on phenotype-genotype relationships for crop production (Slafer et al., 2014; Ferrante et al., 2017).

In the past decade, image processing has been increasingly used in analyzing and extracting phenotypic parameters. Features that include color, texture, shape, and edge are fused in the classifier to detect wheat spikes using traditional image processing methods. Mirnezami et al. (2020) compared automated and semiautomated soybean trichome counting methods, which used thresholding and graph algorithms based on color and shape features. They achieved approximately 90% accuracy using semiautomated annotation, which outperformed manual counting. Kulkarni and Patil (2012) employed the Gabor filter to detect plant diseases by extracting typical plant features from red–green–blue (RGB) images, including texture, edge, and color for plant disease segmentation. Then, the features were used to train the artificial neural network, and the accuracy reached 91%. Sun et al. (2019) applied a region growing algorithm with a double threshold integrating spatial and color features to segment cotton bolls and developed an algorithm based on geometric features to count cotton bolls. The counting accuracy was 84.6%, and the F1 score was approximately 98%. The panicle segmentation method extracted the color and texture of the panicles to realize (semi) automatic counting of wheat spikes (Cointault et al., 2008). Fernandez-Gallego et al. (2018) presented an automatic spike-counting method to calculate the number of spikes based on color images taken under natural conditions. Additionally, the local peaks are segmented and counted by the color features and the Find Maxima. The results showed that the accuracy of wheat spikes counting is 90%, and the standard deviation is 5%. Although most of these studies achieved good results, there were still problems. They have used the traditional image processing method and therefore require manual screening of features. This limitation hinders the popularization and application of the algorithm in more complex problems. The wheat spike detection and counting is still a very challenging task.

Deep learning performs exceptionally well in detection and classification tasks. A series of novel deep learning models have been developed, such as region-based convolutional neural network (R-CNN), Fast R-CNN, Faster R-CNN, fully convolutional one-stage object detector (Fcos), You Only Look Once (Redmon et al., 2016; Redmon and Farhadi, 2017) version 3 (Yolov3), You Only Look Once version4 (Yolov4), You Only Look Once version 5 (Yolov5), RetinaNet, and single-shot multiBox detector (SSD) (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Liu et al., 2016; Lin et al., 2017a; Redmon and Farhadi, 2018; Tian et al., 2019, Bochkovskiy et al., 2020; Jocher et al., 2020), which are ready to be used in phenotyping applications. Backbone network and feature pyramid network (FPN) (Lin et al., 2017b) are the two main components of an object detection framework. The backbone network conducts feature extraction, whereas FPN conducts feature fusion. As a result, advancements in the backbone network and FPN directly impact the performance of the object detection network. He et al. (2016) proposed residual network (ResNet), introducing residual blocks and realizing across layer information transmission through shortcut connections resulting in improved optimization. After that, many studies designed various modules to strengthen the ability of network feature extraction. For example, selective kernel (SK) block (Li et al., 2019), squeeze-and-excitation (SE) block (Hu et al., 2018), non-local block (Wang et al., 2018), convolutional block attention module (CBAM) (Woo et al., 2018), split attention block (Zhang et al., 2020), etc. FPN fuses multiscale features extracted through deep convolutional networks. Tan et al. (2020) proposed a simple and efficient feature pyramid structure to address the top-down architecture of FPN, which is called a bidirectional feature pyramid network (BiFPN). It allows top-down and bottom-up multi-scale weighted feature fusion.

Wheat spike image sets, such as ACID (Pound et al., 2017) and SPIKE (Hasan et al., 2018) were used in many studies and they achieved good deep learning model training results (Alkhudaydi and Zhou, 2019; Madec et al., 2019; Yang et al., 2019). Misra et al. (2021) developed an online platform "Web-spikeSegNet" that uses deep learning methods to segment wheat spike images taken under laboratory environment conditions. It can achieve 99.59% segmentation accuracy. Zhao et al. (2021) proposed an improved Yolov5 network by adding a microscale detection layer, setting prior anchor boxes, and adapting the confidence loss. These improvement points solve spike error detection and miss detection caused by occlusion conditions in UAV images. These studies used deep learning methods to overcome the disadvantages of traditional image processing methods that require manual feature design. However, the datasets used in these studies are relatively homogeneous in terms of wheat spike collection environments and varieties. Most wheat spike datasets are limited in terms of genotype, geographic area, and observational condition. Therefore, the research requires a richer dataset and the ability to overcome the detection of wheat spikes in complex environments. The Global Wheat Detection (GWHD) dataset (David et al., 2020) was a standard image set collected by several research institutions, which was considered by many scholars as a new challenge for wheat spike detection. Bhagat et al. (2021) proposed a novel WheatNet-Lite network, which was solved the dense and overlapping wheat spikes. The network was validated on GWHD, SPIKE, and ACID datasets. The mAP50 values were 91.32, 86.10, and 76.32%, respectively. Li et al. (2021) also investigated the GWHD dataset. They trained RetinaNet models using migration learning. The images of wheat at the filling stage and the maturity stage from the GWHD dataset were used for regression analysis of count results. The R^2 was 0.9722. Wang et al. (2021) proposed an occlusion robust wheat spike counting algorithm based on EfficientDet-D0 with the CBAM attention module. It was the network that focused more on small wheat spikes with the counting accuracy which was 94% and the false detection rate was 5.8% on the GWHD dataset. The new models in these studies were proposed to solve

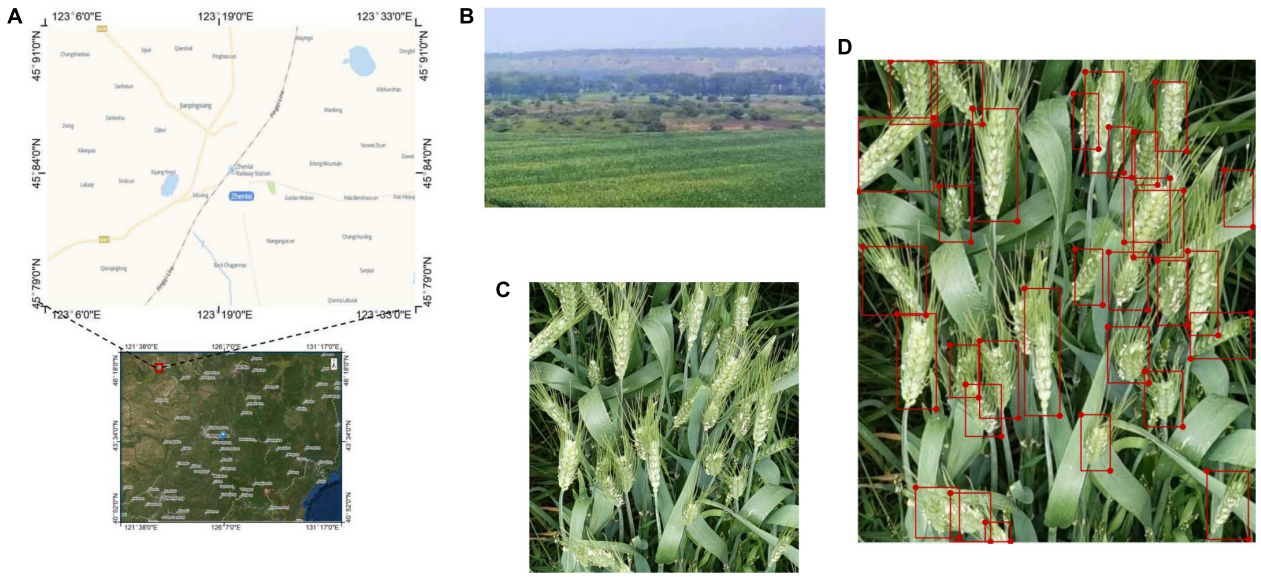

FIGURE 1 | Data acquisition and labeling process: **(A)** map of data collection place; **(B)** the WSD images collection scene; **(C)** an example WSD image; **(D)** annotation of a WSD images.

the wheat spike images occlusion problem. However, it is not only the occlusion images of wheat spikes that are difficult to recognize in the field, but also difficult to recognize wheat spike images with dim lighting and complex environmental backgrounds. Therefore, there is still room for continued improvement in wheat spike detection and counting. In this study, we used the GWHD dataset supplemented with wheat-wheatgrass spike detection (WSD) images, where WSD was collected from trials. There is one variety in our dataset, Jilin wheat-wheatgrass No. 37. Because of its excellent quality, wheat-wheatgrass has been crowned as a geographical landmark product of Jilin Province. The spike of wheat-wheatgrass No. 37 is rectangular in shape, and the spike length is usually 10–12 cm. The wheat spikes have white hulls and are awned but without hairs. WSD images added diversity to the GWHD to train spike detection models.

In this study, SpikeRetinaNet was trained to detect wheat spikes based on the RetinaNet network structure of a one-stage detector, which kept the one-stage detector's speed while improving detection accuracy. In the dataset, it is difficult to distinguish wheat spikes because of light, shadows, color, and shape similarity. To solve the problems, the focal loss function was introduced into the structure of RetinaNet to reduce the influence of background during wheat spike detection tasks (Lin et al., 2017a). Meanwhile, we introduced the BiFPN (Tan et al., 2020) and double SA (DSA) (split attention block and spatial attention block) into the backbone network to realize fine-grained feature extraction and representation across feature map groups and strengthen the fusion of global information and local information. By proposing BiFPN, it introduces learnable weights to learn the importance of different input features and repeatedly applies top-down and bottom-up multiscale feature fusion. Because different input features have different resolutions, their contribution to the fused fine-grained features is different. Meanwhile, introducing DSA into the backbone realizes the interaction between feature map channels and receptive field regions. In this way, fine-grained discriminant feature of detecting wheat spikes, such as the shape, texture, and color, can be better extracted and represented. The cluster growth of wheat spikes makes it difficult to distinguish between multiple wheat spikes or multinode parts of wheat spikes because wheat spikes occlude each other. In the previous work, non-maximum suppression is an integral part of the object detection pipeline which is used to filter the detection candidate boxes. The detection box with the maximum score is selected and all other detection boxes with a significant overlap (using a predefined threshold) are suppressed. To this end, we introduced soft non-maximum suppression (Neubeck and Van Gool, 2006) (Soft-NMS) (Bodla et al., 2017), an algorithm that decays the detection scores of all other objects as a continuous function of their overlap, to solve the problem of missed detection caused by mutual occlusion.

MATERIALS AND METHODS

Image Data Acquisition

The original GWHD dataset included 4,700 high-definition color images of wheat from multiple genotypes. There were a total of 190,000 wheat spikes annotated. Wheat spikes in the image were labeled interactively by delimiting bounding boxes that contained all spike's pixels using web-based labeling (Brooks, 2019). Seven categories that contain 3,373 images and 147,793 labeled spikes from Europe and North America were used in this article. The seven categories are Arvalis_1, Aralis_2, Aralis_3, INRAE_1, USask_1, RRes_1, and ETHZ_1. They are collected between 2016 and 2019. They were acquired over experiments

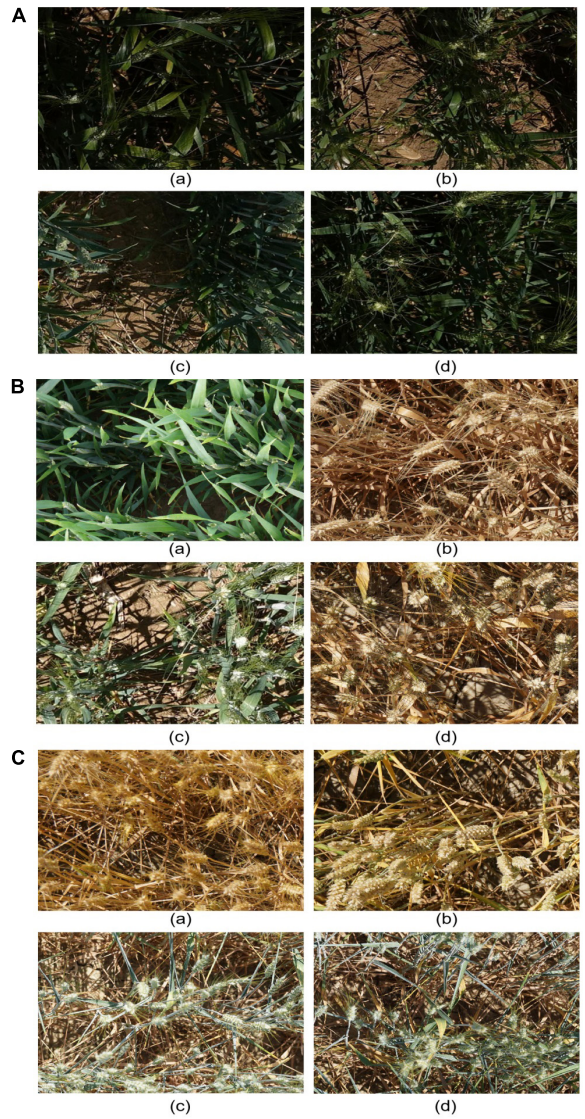

FIGURE 2 | Sample images taken under different complex conditions. **(A)** Dim illumination conditions. **(B)** The complex environment illumination conditions. **(C)** Overlapping occlusion of wheat spikes.

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, and INRAE_1) and high sowing density (RRes_1 and ETHZ_1). The GWHD dataset covers a range of pedoclimatic conditions including very productive contexts, 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 and Arvalis_2). In the case of Arvalis_1 and Arvalis_2, the experiments were designed to compare irrigated and water-stressed environments. An average of 44 spikes was present in each image, with a range of 15–70 real spikes per image. The WSD images that contain 210 high-definition color images and 6,123 annotations were used in this study to supplement the experimental data as well. All images were collected from Chengkai Cooperative, Nangangzi Village, Zhenlai Town, Baicheng City, Jilin Province, China (45.83 N, 123.21 E) from May to July 2020 using a Canon 11 EOS 80D digital camera. Images were captured at the height of 30–70 cm above the wheat canopy and at various tilt angles. The resolution of the WSD images was 3,456 × 4,408 pixels. All images were stored in JPG format according to the RGB color standard. Then, the collected images were labeled by the LabelImg Tool (LabelImg, 2015). The overall process is shown in **Figure 1**.

Formation of the Mixed Dataset

The original image set was first normalized to obtain a total of 3,583 images of 1,024 × 1,024 pixels due to the limited computing power of laboratory equipment. The diversity and complexity of the mixed dataset brought great difficulties to the method in detecting and counting. Three image categories were the most difficult to identify: (1) images with low illumination, (2) complex environment, and (3) overlapping objects. For example, it is difficult to distinguish wheat spikes in the evening due to dim light and complicated shadows (**Figure 2A**). When wheat plants are young (**Figure 2B(a)**), their spikes are small and as green as the leaves. Wheat spikes (**Figure 2B(b,d)**) and stems are similar in color too, and there is a mutual occlusion phenomenon, which can easily confuse analysis. **Figure 2B(c,d)** is sparsely planted, with a visible soil background, and the distribution of shadows is mixed by light. The cluster growth of wheat spikes in **Figures 2C(a,b)** makes it difficult to distinguish between multiple wheat spikes or multinode parts of wheat spikes. In **Figures 2C(c,d)**, wheat spikes occlude each other, which make it difficult to mark.

Our next step filtered out some inappropriate bounding boxes [the boundary box is too large (box areas >200,000) or too small (box areas <50)] from the dataset before putting the images into the model to make it more accurate and clean. Then, we used online augmentation techniques, such as horizontal and vertical flips, rotations and resizing, and augmenter and normalizer to enhance the image. This method has the advantage of not requiring the augmented data to be synthesized, which saves data storage space and provides high flexibility. Among the 3,583 wheat images collected, 70% of each category in the mixed dataset was extracted as the training dataset, 20% of images were extracted as the validation set, and 10% of images were extracted as the test set.

Overall Design of the SpikeRetinaNet

Figure 3 depicts the specific process of our proposed SpikeRetinaNet. First, the image features are extracted through the convolution layer. Then, the extracted feature sets are grouped and convoluted to calculate the weight of the feature channel and then performing a weighting operation on the obtained weights and feature sets. Second, we perform AdaptiveAvgPool2d and AdaptiveMaxPool2d on the results obtained. We then use the sum of the pooling results to calculate the weight value through the Sigmoid function and then performing another weighting operation on the weight value and feature set to get the result of spatial attention. Third,

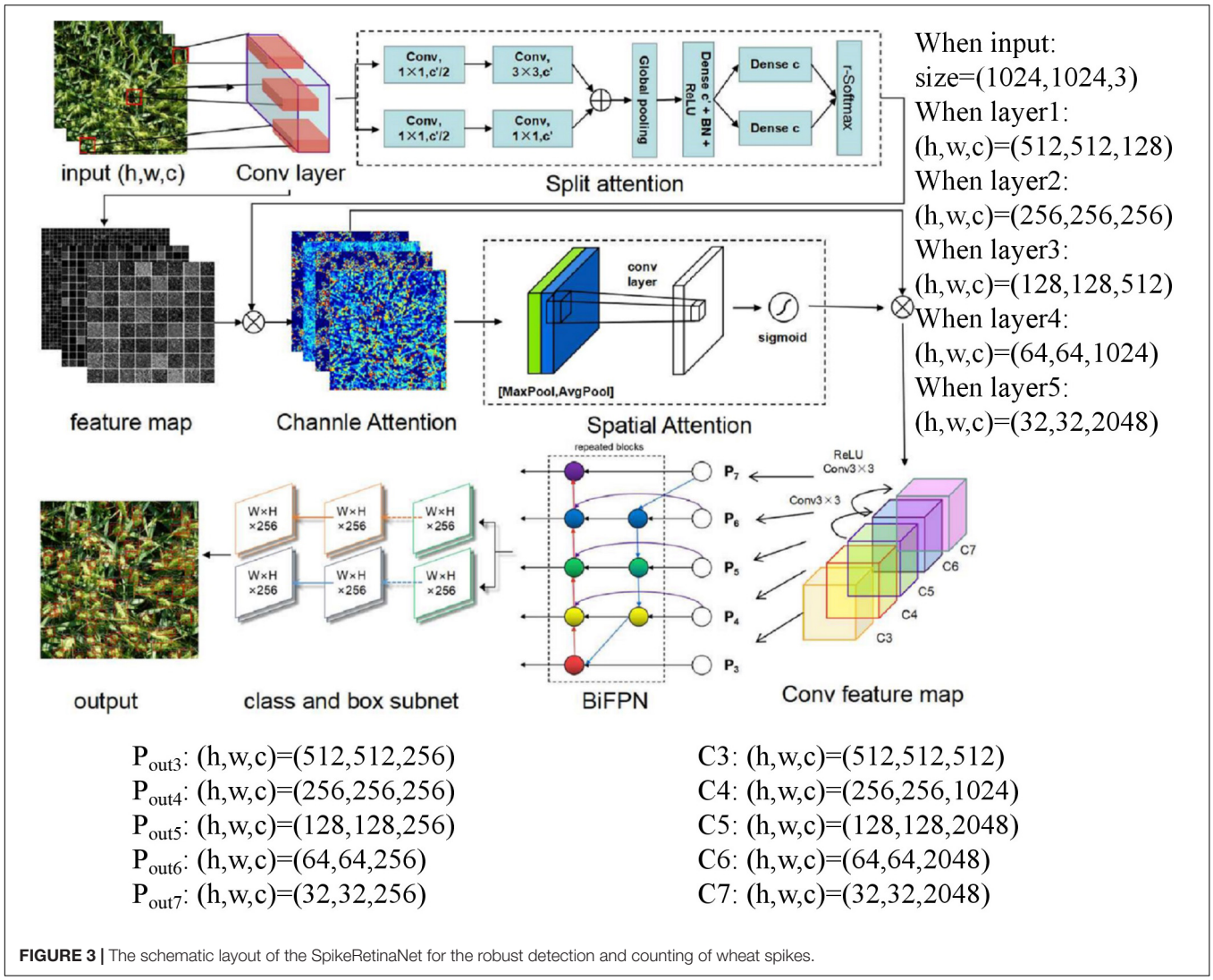

FIGURE 3 | The schematic layout of the SpikeRetinaNet for the robust detection and counting of wheat spikes.

SpikeRetinaNet employs five levels of feature pyramids. P3, P4, and P5 are calculated by top-down and lateral connections of the corresponding backbone network's C3, C4, and C5 layers (architectures for ImageNet (Krizhevsky et al., 2012) are divided into C1–C5), respectively. P6 is obtained by upsampling based on C5, and ReLU obtains P7 based on P6. The output is obtained by weighted bidirectional calculation of P3–P7. Finally, the results of each layer of the FPN are input into two subnetworks of classification and regression, respectively, to get the final output image.

RetinaNet

RetinaNet is an object detector that consists of a backbone network and two task-specific subnetworks. Among them, the backbone networks include a convolutional neural network to extract information from the image and the FPN enhancing the feature information with top-down and lateral connections. The two subnets use convolution to classify and regress from bounding boxes to real object boxes.

The core of RetinaNet is focal loss. It simply and efficiently solves the category imbalance faced by the one-stage detector, which improves the classification precision of the one-stage detector. RetinaNet was proposed to reshape the standard crossentropy loss to focal loss to deal with the category imbalance. It downweighs simple samples so that even if the number of samples is large, their contribution to the total loss is small. The focal loss formula is as follows Eq. 1.

$$FL(p_t) = -\alpha t(1-pt)^{\gamma}\log(pt)$$
$$pt = \begin{cases} p & if\ \ \gamma = 1 \\ 1-p & otherwise \end{cases} \quad (1)$$

The weighting factor $\alpha \in [0, 1]$ is the parameter for class 1, and $1-\alpha$ for class -1, α maybe set by inverse class frequency or treated as a hyperparameter set by cross-validation. Though α balances the weight values of positive or negative examples, it does not differentiate between easy or hard examples. So, the modulating factor $(1-pt)^{\gamma}$ is introduced with a tunable focusing parameter and $\gamma \geq 0$ and pt is the class probability score. The proposed adjustment factor reduces the loss weights ratio from simple examples and quickly focuses on hard examples. It is suitable for difficult distinguishing between foreground and

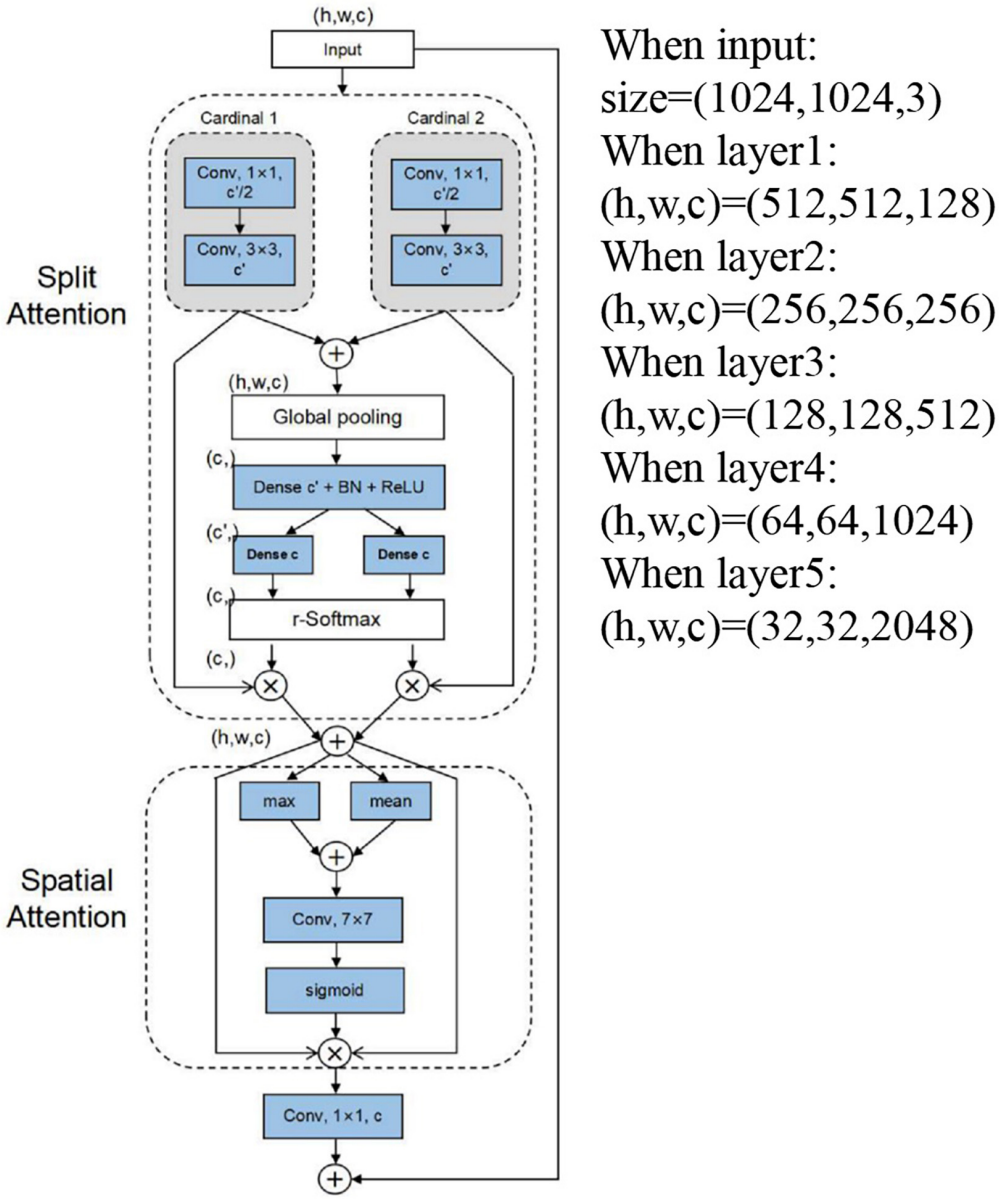

FIGURE 4 | Schematic layout outline of ResNet-DSA.

background, such as many negative examples in the process of wheat spike detection. Therefore, when discussing dense object detection (such as our mixed dataset), RetinaNet is the best choice for speed and accuracy.

Selection of the Feature Learning Network

The design of the feature learning network is very important. We add the DSA (double SA, split attention block, and spatial attention block) to the backbone network of RetinaNet to enable the feature mapping attention between different feature mapping groups and emphasize the spatial location information. Further detailed description in **Figure 4** divides the features into two groups (V_1 and V_2) for 1×1 convolution followed by a 3×3 convolution. The attention weight is parameterized using two fully connected layers with ReLU activation. We aggregate channel information of a feature map using two pooling operations (maxpool and avgpool), generating two 2D maps. Then, we connect them and convolute them through standard convolution operation to form our 2D spatial attention maps. Finally, if the input and output feature maps have the same size, the final output Y of our DSA is produced using a shortcut connection: $Y = V + X$ ($V = Cancat\{V_1, V_2\}$). For blocks with a stride, an appropriate transformation $T(X)$ is applied to the shortcut connection to align the output shapes: $Y = V + T(X)$. The specific shape is depicted in the note of **Figure 4**, where the feature maps become smaller and

TABLE 1 | Architectures for ImageNet.

Layer name	Output size	SpikeRetinaNet	RetinaNet
Conv1	112 × 112	[3 × 3] × 3, 64, *stride* 2	[7 × 7], 64, *stride* 2
Conv2_x	56 × 56	3 × 3 max*pool*, *stride* 2	3 × 3 max*pool*, *stride* 2
		$\begin{bmatrix} 1\times1,\ 128 \\ \textit{Split Attention},\ C=2, G=64 \\ \textit{Spatial Attention} \\ 1\times1,\ 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1,\ 64 \\ 3\times3,\ 64 \\ 1\times1,\ 256 \end{bmatrix} \times 3$
Conv3_x	28 × 28	$\begin{bmatrix} 1\times1,\ 256 \\ \textit{Split Attention},\ C=2, G=128 \\ \textit{Spatial Attention} \\ 1\times1,\ 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1,\ 128 \\ 3\times3,\ 128 \\ 1\times1,\ 512 \end{bmatrix} \times 4$
Conv4_x	14 × 14	$\begin{bmatrix} 1\times1,\ 512 \\ \textit{Split Attention},\ C=2, G=256 \\ \textit{Spatial Attention} \\ 1\times1,\ 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1,\ 256 \\ 3\times3,\ 256 \\ 1\times1,\ 1024 \end{bmatrix} \times 23$
Conv5_x	7 × 7	$\begin{bmatrix} 1\times1,\ 1024 \\ \textit{Split Attention},\ C=2, G=512 \\ \textit{Spatial Attention} \\ 1\times1,\ 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1,\ 512 \\ 3\times3,\ 512 \\ 1\times1,\ 2048 \end{bmatrix} \times 4$
P3_out	80 × 80	*Conv*3 → 1 × 1	*Conv*3 → 1 × 1
P4_out	40 × 40	*Conv*4 → 1 × 1	*Conv*4 → 1 × 1
P5_out	20 × 20	*Conv*5 → 1 × 1	*Conv*5 → 1 × 1
P6_out	10 × 10	*Conv*5 → 3 × 3	*Conv*5 → 3 × 3
P7_out	5 × 5	$\textit{Conv}5 \rightarrow \begin{bmatrix} 3\times3 \\ \textit{ReLU} \\ 3\times3 \end{bmatrix}$	$\textit{Conv}5 \rightarrow \begin{bmatrix} 3\times3 \\ \textit{ReLU} \\ 3\times3 \end{bmatrix}$
Regression	Px.size()	$\begin{bmatrix} 3\times3,\ 256 \\ \textit{ReLU} \end{bmatrix} \times 4$ 3 × 3, 36	$\begin{bmatrix} 3\times3,\ 256 \\ \textit{ReLU} \end{bmatrix} \times 4$ 3 × 3, 36
Classification	Px.size()	$\begin{bmatrix} 3\times3,\ 256 \\ \textit{ReLU} \end{bmatrix} \times 4$ 3 × 3, 63 *Sigmoid*	$\begin{bmatrix} 3\times3,\ 256 \\ \textit{ReLU} \end{bmatrix} \times 4$ 3 × 3, 63 *Sigmoid*

Building blocks are shown in brackets, with the numbers of blocks stacked. C is the number of groups, and G is the number of channels per group.

the channels become more numerous as the network depth deepens. The backbone network has better and more accurate feature extraction capabilities than ResNet. Therefore, we can extract more detailed features for the spike of wheat detection. For the problem of dim light and complex environment background in the mixed datasets, we can apply the DSA attention module to emphasize the characteristics of wheat spikes. Similarly, suppose the wheat spikes in the data set are similar to the background. In that case, we can also use the attention block to emphasize the useful features and suppress the useless features.

Design of the Feature Pyramid Network Backbone

We use BiFPN of FPN to enhance feature fusion. BiFPN can realize fast bidirectional cross-scale connections and weighted feature fusion. Among them, multiscale feature fusion is to be carried out using different levels and different resolutions of the input. This produces a list of multiscale features $\vec{P^{in}} = (P^{in}_{l_1}, P^{in}_{l_2}, \ldots)$, which $P^{in}_{l_i}$ represents the feature at a level l_i. BiFPN requires P^{in}_3 to P^{in}_7 level inputs for aggregate features. The traditional output calculation of FPN is shown in Eq. 2, where *Resize* is the upsampling or downsampling operators to adjust the image size and *Conv* is a convolutional operator.

$$\begin{aligned} P^{out}_7 &= Conv(P^{in}_7) \\ P^{out}_6 &= Conv(P^{in}_6 + \mathrm{Resize}(\mathrm{P}^{\mathrm{out}}_7)) \\ &\cdots \\ P^{out}_3 &= Conv(P^{in}_3 + \mathrm{Resize}(\mathrm{P}^{\mathrm{out}}_4)) \end{aligned} \tag{2}$$

TABLE 2 | Mean average precision (mAP), frames per second (FPS), root mean square error (RMSE), and root mean square percentage error (RMSPE) of RetinaNet in detecting wheat spikes.

Method	Datasets	mAP50	mAP75	FPS	RMSE	RMSPE	Counting Acc
RetinaNet	Mixed	0.8703	0.4701	35	2.63	0.08	0.8984
RetinaNet-DSA	Mixed	0.9143	0.4842	30	–	–	0.9122
RetinaNet-DSA-BiFPN	Mixed	0.9243	0.4942	25	–	–	0.9206
Our method	**Mixed**	**0.9262**	**0.5023**	**22**	**1.96**	**0.06**	**0.9228**

The results of our method are highlighted in bold.

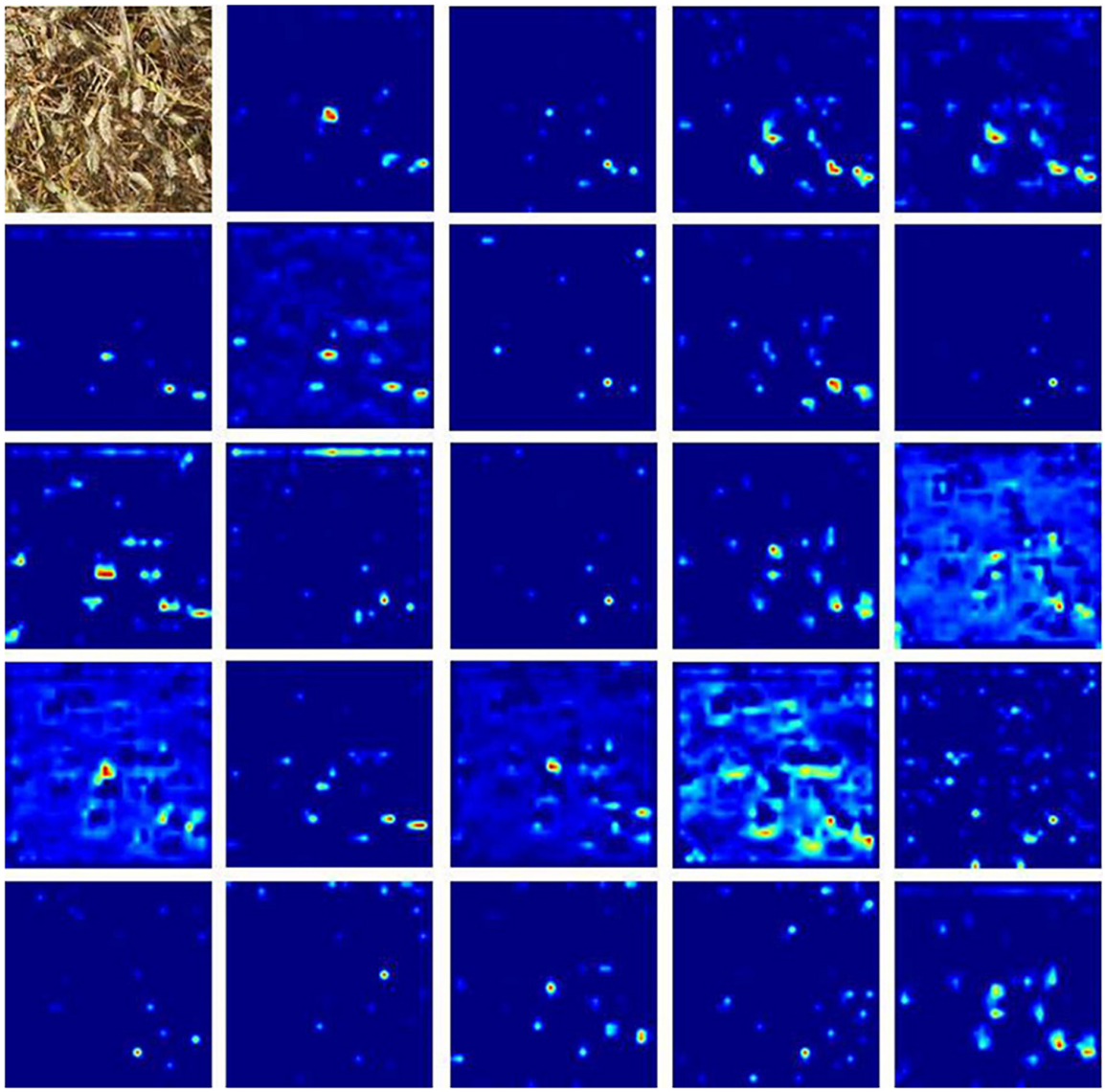

FIGURE 5 | The first image is the original image, the other images are the class activation mapping (CAM) for the different feature channels of our method.

Therefore, BiFPN adds an extra bottom-up path aggregation network to solve the problem that conventional FPN only has top-down unidirectional information flows. Besides, the bidirectional network is simplified by removing the node with only one input channel to integrate more features without increasing much cost. Therefore, we represent the fused feature at level six for BiFPN shown in Eq. 3:

$$\begin{aligned} P_6^{td} &= Conv\left(\frac{\omega_1 \cdot P_6^{in} + \omega_2 \cdot \text{Resize}(P_7^{in})}{\omega_1 + \omega_2 + \varepsilon}\right) \\ P_6^{\text{out}} &= Conv\left(\frac{\omega_1^{'} \cdot P_6^{in} + \omega_2^{'} \cdot P_6^{td} + \omega_2^{'} \cdot \text{Resize}(P_5^{\text{out}})}{\omega_1^{'} + \omega_2^{'} + \omega_3^{'} + \varepsilon}\right) \end{aligned} \tag{3}$$

P_6^{td} is the median feature at level six on the top-down pathway and P_6^{out} is the output feature at level six on the bottom-up pathway. The bidirectional fusion of BiFPN deepens the degree of feature fusion. So, in the mixed dataset, images with complex environment backgrounds can use deep, low-resolution, and high semantic features to distinguish wheat spikes and background. As a result, more overlapping wheat spikes can be retained. Meanwhile, shallow, high-resolution features could provide more accurate location information. It can also locate the problem of wheat spikes occlusion better.

Soft Non-maximum Suppression

Soft non-maximum suppression was introduced to obtain consistent improvements for the selection of candidate boxes. Soft-NMS suppresses overlapping boxes with a non-maximum value and sets the attenuation function for near boxes based on the overlapping boxes' size instead of setting its score to zero.

Intuitively, if the crossarea between the bounding box and M is higher than the threshold, its score should be reduced. If its overlap is lower than the threshold, it keeps the detection score unchanged. The calculation formula is shown in Eq. 4, where S_i is the final score, i is the subscript, M is the box with the highest score in the prediction box set, b_i is the box in the prediction box set B, and N_t is the intersection-over-union (IoU) threshold of M and b_i. The formula Eq. 5 updated the pruning step with the following rule. Under natural conditions, the presence of wheat spikes occlusion is inevitable in wheat spikes data collection. The Soft-NMS can effectively retain the blocked wheat spikes without affecting the selection of the normal calibration box.

$$S_i \begin{cases} S_i, & iou(M, b_i) < N_t \\ S_i(1 - iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \quad (4)$$

$$S_i = S_i e^{-\frac{iou(M,b_i)^2}{\sigma}}, \forall b_i \notin D \quad (5)$$

See **Table 1** for detailed architectures compared the SpikeRetinaNet with the original RetinaNet.

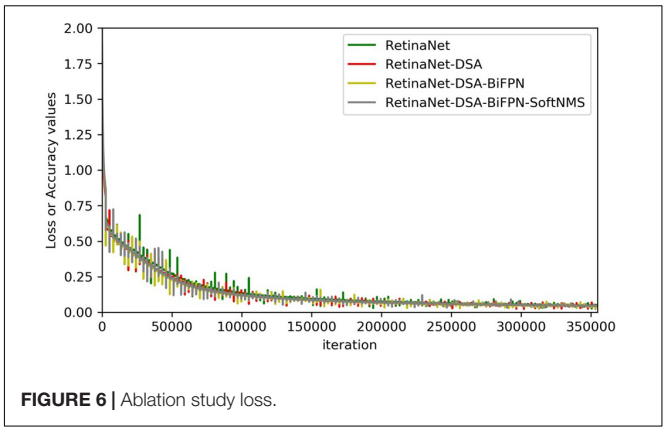

FIGURE 6 | Ablation study loss.

TRAINING THE WHEAT SPIKE DETECTING AND COUNTING MODEL

Computational Hardware and Platform

All processing experiments in this article were carried out by the DELL Precision T7920 Tower deep learning workstation which consisted of an Intel(R) Xeon(R) Gold 5218 CPU with a clock speed of 2.1 GHZ, 62.5 GB DRAM, 503 GB hard disk, and a GeForce RTX 2080 Ti/PCIe/SSE2 graphics card. The operating environment was Ubuntu 18.0.4, Pytorch = 1.7.0, Python 3.7.

Model Training

The SpikeRetinaNet was used in the mixed dataset training process. First, DSA was used to extract the features from the backbone network. Second, BiFPN improved the extracted feature map by adding more expression and multiscale target region data. Finally, two subnetworks with the same structure but no shared parameters used BiFPN feature maps to complete the object classification task and regress the offset from the bounding box to a nearby real object. The Soft-NMS was also used to choose calibration boxes. The parameters of the different layers are described in the note of **Figure 3**. The specific algorithm flow is as follows:

Input: image to be detected D.

Output: Vector C is used for sample categories, and R is used for boundary coordinates.

Step 1: Convolution layer for feature extraction. First, 64 convolution kernels with 7 × 7 stride-2 are used to feature extraction, and then, a maxpooling with 3 × 3 stride-2 is used to get the feature set. Second, all feature maps are divided into 2 splits. Additionally, the split attention is used to calculate the weight of each split, and the weighted feature maps are used as the input of the spatial attention module. Finally, a 1 × 1 Conv is used again to change the number of channels and use skip connection to fuse the original input features of a DSA block (the fusion method is element-wise sum). There are 101 layers as a feature extraction network.

Step 2: FPN, the multiscale features formed in the backbone network, is input into the feature pyramid for enhancement and utilization, and the feature map with stronger expression and multiscale target information is obtained. The backbone network is divided into C1–C5 layers. Add a 1 × 1 Conv on C5, and the upsampling is two times as much to generate the feature map, and then, a ReLU activation function is performed to form P_7^{td}. P_6^{td} is to add a 1 × 1 Conv on C5, and the upsampling is two times as much to generate the feature map and then fuse with P_7^{td}. P_5^{td} is directly mapped from C5 to merge P_6^{td} upsampling. P_4^{td} and P_3^{td} have the same structure as P_5^{td}. P_3^{td} to P_7^{td} is the input of the FPN. P_3^{out} is upsampled by C3 fusion P_4^{td}, P_4^{out} is formed by P_4^{td} and C4 fusion P_3^{td} downsampling. P_5^{out} and P_6^{out} have the same structure as P_4^{out}. P_7^{out} is downsampled and fused by P_6^{td} and P_6^{out}. Finally, a 3 × 3 Conv stride-2 is used for all the layers obtained after fusion to eliminate the aliasing effect of upsampling.

Step 3: The output of each layer of the feature pyramid performs two subnetwork tasks (classification and boxes regression). Each subnetwork uses four layers of 3 × 3 × 256 Conv and then connects to 3 × 3 × KA (K is the number of object classes, A = 9 anchors per level) Conv. In addition, it finally uses Sigmoid activation to the output KA binary predictions at each spatial position.

Step 4: Use a trained model to perform the next decoding process on the top 1,000 boxes with the highest scores on each FPN level. Summarize boxes of all levels, filter boxes with a soft threshold of 0.1, and finally get the final boxes location of the target. The training loss is composed of boxes position information L1 loss and category information Focal-Loss. Considering the extreme imbalance between positive and negative samples when the model is initialized, the bias parameter of the last convolution is initialized.

The specific steps of the training are as follows: due to equipment limitations, a minibatch of four images will be used to train the model. The Optimizer selects Adam, uses Reduce LROnPlateau to dynamically adjust the learning rate, the initial learning rate is 1e-4, and uses all images of the training

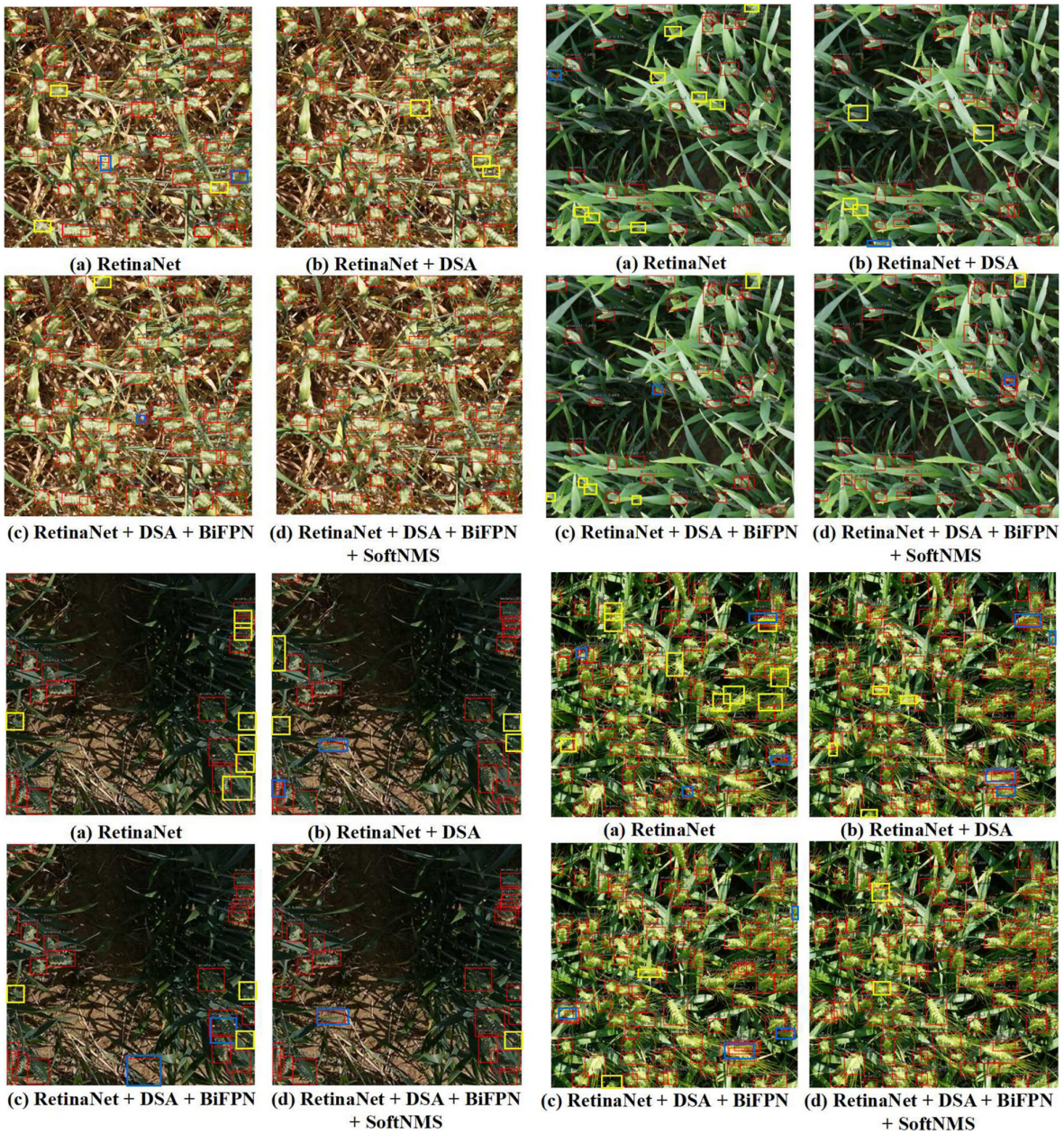

FIGURE 7 | Four different examples, each with the four different methods. **(a)** Results of wheat spike detection with RetinaNet. **(b)** Results of wheat spike detection with RetinaNet + DSA. **(c)** Results of wheat spike detection with RetinaNet + DSA + BiFFPN. **(d)** Results of wheat spike detection with our method. Additionally, yellow boxes are missed spikes, blue boxes are false spikes.

dataset to train 100 epochs to analyze the training process. Additionally, the same platforms are also applied to faster region-based convolutional neural network (Faster-RCNN), YoLov3, YoLov4, YoLov5s, YoLov5m, and SSD, which codes are publicly available for comparison.

Network Evaluations

For this study, all samples were divided into four types according to the IoU between the predicted bounding boxes and the real bounding boxes exceeding a given parameter. True positive (TP) corresponds to the correct predicted bounding boxes. False-positive (FP) corresponds to the erroneously predicted bounding boxes. False-negative (FN) is the marked bounding box that could not be detected. Otherwise, it is a true negative (TN). Eq. 6 precision (P) and Eq. 7 recall (R) are computed.

$$P = \frac{TP}{TP + FP} \tag{6}$$

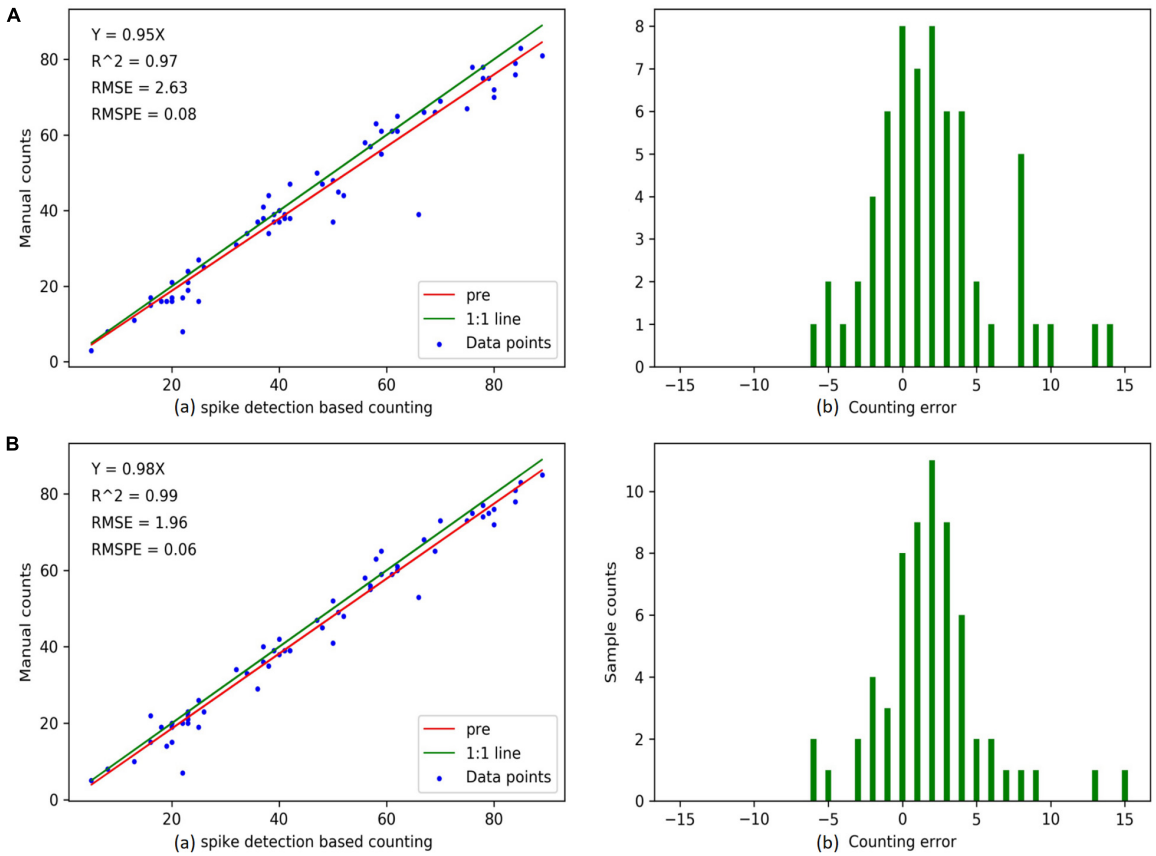

FIGURE 8 | Counting accuracies calculated using "spike detection based counting" **(top row)** and "manual counting" **(bottom row)** strategies, respectively, for individual spikes (a total of 60 data points). The above two images are the results of RetinaNet, and the following two images are the results of our method. **(A(a),B(a))** are linear regression results between the imaging derived and manual counts. **(A(b),B(b))** are the histograms of counting errors. The "pre" in graph "spike detection-based counting" indicates the predicted regression line.

$$R = \frac{TP}{TP + FN} \tag{7}$$

Since the evaluation index mainly focuses on the positive sample, thus to weigh the precision index and the recall index, AP_k (the value k represents the type of wheat spikes) was defined in Eq. 8 as the area under the P_k and R_k curve of the class k. AP is a standard measure to measure the sensitivity of the network to target objects, and it is also an indicator of the overall performance of the network. Additionally, mAP was defined in Eq. 9 as the average precision of the eight classes of wheat spikes. The higher the mAP, the better the detection results of the convolutional neural network for the object detection, and the average detection time is also calculated to evaluate the performance of the model.

$$AP_k = \int_0^1 P(R_k) dR_k \tag{8}$$

$$mAP = \frac{1}{8} \sum_{k=1}^{8} AP_k \tag{9}$$

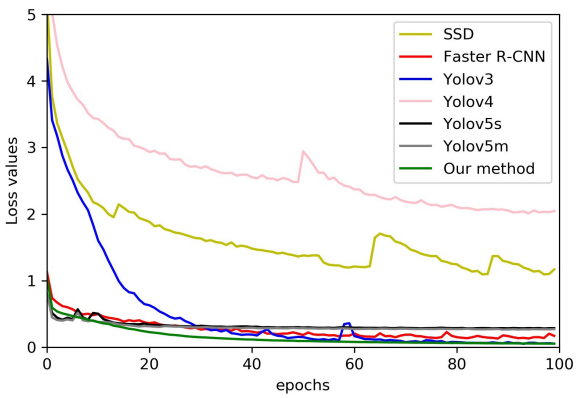

FIGURE 9 | Our method comparing against the state-of-the-art loss.

Two other metrics were proposed to evaluate the performance of spikes counting: root mean square error (RMSE) as Eq. 10 and root mean square percentage error (RMSPE) as Eq. 11, which

TABLE 3 | Mean average precision (mAP), frames per second (FPS), root mean square error (RMSE), and root mean square percentage error (RMSPE) of SSD, Yolov3, Yolov4, Yolov5s, Yolov5m, Faster R-CNN, and our method in detecting wheat spikes.

Method	Backbone	Datasets	mAP50	mAP75	FPS	RMSE	RMSPE	Counting Acc
SSD	VGG	Mixed	0.4356	0.1652	60	10.30	0.26	0.4841
Yolov3	DarkNet53 + FPN	Mixed	0.8983	0.4832	50	2.56	0.08	0.8991
Yolov4	CSPDarkNet53 + PANet	Mixed	0.9127	0.4902	52	2.13	0.14	0.9095
Yolov5s	CSPDarkNet53 + PANet	Mixed	0.9272	0.5128	60	1.71	0.12	0.9302
Yolov5m	CSPDarkNet53 + PANet	Mixed	0.9312	0.5217	50	1.53	0.06	0.9330
Faster-RCNN	ResNet101 + FPN	Mixed	0.8536	0.4956	10	3.14	0.07	0.8805
Our method	ResNet101 + BiFPN	**Mixed**	**0.9262**	**0.5023**	**22**	**1.96**	**0.06**	**0.9288**

The results of our method are highlighted in bold.

N_p is the predicted value of wheat spikes and N_g is the actual value of wheat spikes. The number of spikes detected by the model and the number of spikes counted manually were analyzed by simple linear regression. The coefficient of determination R^2 was calculated to assess the effectiveness of using one variable to predict the other.

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(N_p - N_g)^2} \tag{10}$$

$$RMSPE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left|\frac{(N_p - N_g)}{N_p}\right|^2} \tag{11}$$

RESULTS

Ablation Study

Evaluation of the SpikeRetinaNet

In this subsection, we empirically show the effectiveness of our design choice. As shown in **Table 2**, the results indicate that the effect of added DSA blocks to RetinaNet is better than the original network, and the mAP is increased by 4.40%. The RetinaNet-BiFPN is better than RetinaNet-FPN, and the mAP is increased by 1%. Therefore, our model can improve the mAP of RetinaNet by about 5.40%. The class activation mapping (Selvaraju et al., 2017) of our model is shown in **Figure 5**. The CAM uses the gradient information from the feature map from the P7 layer of the BiFPN to understand the importance of each feature point to the target decision. The thermodynamic features of different colors reveal the "attractiveness" of the regional network. Among them, the red area represents the most significant influence on the network. As the color changes from red to yellow and finally to blue, it means that the influence has decreased. So, in **Figure 5**, these 24 images represent the visualization result of the 24 feature channels (partial feature channel of P7 layer of BiFPN), thus reflecting our method can focus on wheat spike features in the complex environment. The results show that our backbone has a better capability of feature extraction. Finally, we improve the NMS parts, using Soft-NMS to select candidate boxes, and the performance is improved by 5.59%. The network complexity of our method is increased, so the FPS is reduced from 35 to 22, the increase of time is not much, and the performance is improved significantly. As shown in **Figure 6**, the convergence rate of the loss value is similar in the self-verification comparison experiment, but the fluctuation of RetinaNet is the largest, and our method is the most stable.

Counting Strategy

After detection, 60 images of three categories (low illumination, complex environment background, and overlapping occlusion) are selected for counting, with a total of 1,448 wheat spikes. The counting result of our method is 1,345, and the counting accuracy is 92.88%. The counting result of RetinaNet is 1,301, and the counting accuracy is 89.84%. So, our method has improved by 3.04%. The above experiments indicated that our method could effectively overcome the three kinds of difficult recognition images to improve the precision of spike detection. As can be seen from the following four images (the above two images show a complex environment background, and the next two are low illumination and overlapping occlusion), the counting results of four different networks in the same image are inconsistent (**Figure 7**). Among them, yellow is missing spikes and blue is false spikes. The real counting result is 211, the total counting result of the RetinaNet is 193, and the RetinaNet-DSA-BiFPN result is 205. RetinaNet-DSA-BiFPN [**Figure 7(c)**] can detect wheat spikes that cannot be detected in RetinaNet [**Figure 7(a)**], which indicates that the increased fusion channel makes the fusion information more useful. Finally, the total number of our method [**Figure 7(d)**] is 207. The four images show that the missed boxes of our method are lower than those of other models. This is because Soft-NMS reduces the score of boxes with high IoU rather than directly filtering them out, thus allowing the correct boxes to be retained. The results show that our method improves detection accuracy by 6.63% in images with high detection difficulty.

For the detection results of 60 images, a comparison between the "RetinaNet" and the "our method" is performed (**Figure 8**). The regression slope of "our method" is higher than that of "RetinaNet." In addition, it has a higher correlation, lower RMSE and RMSPE (the RMSE and the RMSPE of our method are 1.96 and 0.06, the RMSE and the RMSPE of RetinaNet are 2.63 and 0.08), which indicates that the counting result of our method (**Figure 8B**) is better than that of RetinaNet (**Figure 8A**). At the

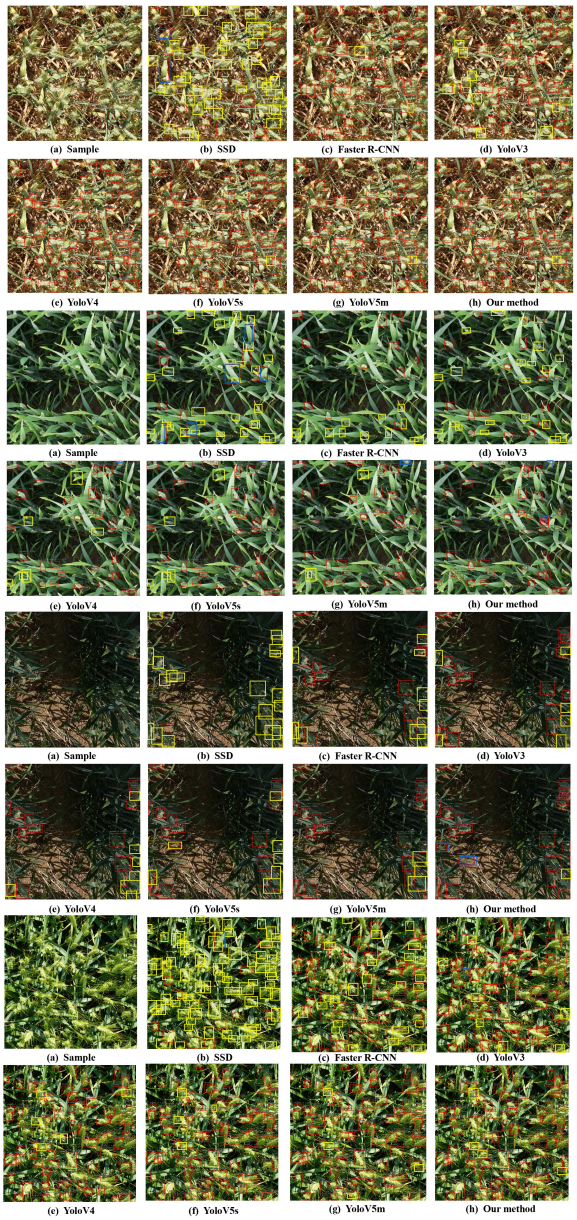

FIGURE 10 | Four different examples, each with the seven different methods: **(a)** sample of images; **(b)** results of wheat spike detection with SSD; **(c)** results of wheat spike detection with Faster R-CNN; **(d)** results of wheat spike detection with Yolov3; **(e)** results of wheat spike detection with Yolov4; **(f)** results of wheat spike detection with Yolov5s; **(g)** results of wheat spike detection with Yolov5m; **(h)** results of wheat spike detection with our method. Additionally, yellow boxes are missed spikes, and blue boxes are false spikes.

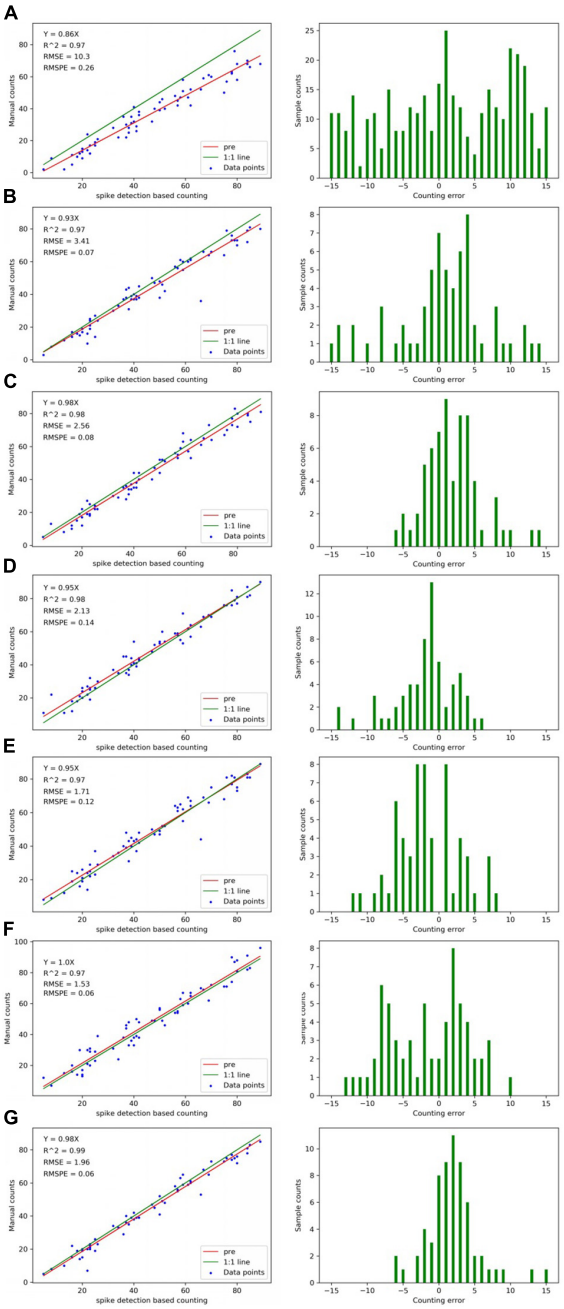

FIGURE 11 | Counting accuracies calculated using "spike detection based counting" (top row) and "manual counting" (bottom row) strategies, respectively, for individual spikes (a total of 60 data points). The **(A)** is SSD, **(B)** is Faster R-CNN, **(C)** is Yolov3, **(D)** is Yolov4, **(E)** is Yolov5s, **(F)** is Yolov5m, and **(G)** is our method. The left is linear regression results between the imaging derived and manual count. The right is the histograms of counting error. The "pre" in graph "spike detection-based counting" indicates the predicted regression line.

same time, in our method, the counting error is concentrated between ±5, which is better than RetinaNet.

Comparing Against the State-of-the-Art Detectors

With mainstream object detection, the one-stage detector used Yolov3, Yolov4, Yolov5s, Yolov5m, and SSD, and the two-stage detector used Faster-RCNN. RetinaNet model is different from the five improved ideas but also has a good detection effect. **Figure 9** shows 100 epoch performances of all models, our

method, Yolov3, Yolov4, Yolov5s, Yolov5m, SSD, and Faster-RCNN. The convergence speed of the loss value of our method is faster than Yolov3, Yolov4, Yolov5s, Yolov5m, SSD, and Faster-RCNN. The final loss of our method is 0.05, Yolov3 is 0.07, SSD is 1.51, Faster-RCNN is 0.15, Yolov4 is 2.04, Yolov5s is 0.28, and Yolov5m is 0.27. Because the one-stage detector does not deal with the detection frame, the initial value of Yolov3 and SSD loss function is greater. As a result of the imbalance between positive and negative examples, the initial value and overall trend pair differ from the other models. Because of the two-stage detector's special RPN network, the convergence speed of Faster-RCNN is slower than that of RetinaNet. Due to our model improving on the loss function, the convergence speed of our model is comparable to Yolov5 and better than the other models. Finally, the mAP value also shows that our method has achieved good experimental results (**Table 3**). The mAP value of our method is 2.79% higher than Yolov3, 1.35% higher than Yolov4, comparable to Yolov5, 7.26% higher than Faster-RCNN, and 49.06% higher than SSD.

After detection, 60 images of three categories (low illumination, complex environment background, and overlapping occlusion) are selected for counting, with a total of 1,448 wheat spikes. The counting result of the Faster-RCNN is 1275, SSD is 701, Yolov3 is 1302, Yolov4 is 1317, Yolov5s is 1347, Yolov5m is 1351, and the counting result of our method is 1345. The counting accuracy of our method is 92.88%, Faster R-CNN is 88.05%, Yolov3 is 89.91%, Yolov4 is 90.96%, Yolov5s is 93.02%, Yolov5m is 93.30%, and SSD is 48.41%. The counting accuracy of our method is 4.83% higher than that of Faster-RCNN, 2.97% higher than that of Yolov3, 1.92% higher than that of Yolov4, and 44.47% higher than that of SSD, and comparable to YoLov5. The above experiments show that our method can effectively overcome the three kinds of difficult recognition images to improve the accuracy of spike detection. As can be seen from the following four images (the above two images show a complex environment background, and the next two are low illumination and overlapping occlusion), the counting results of seven different networks in the same image are inconsistent (**Figure 10**). Among them, yellow is missed spikes, and blue is false spikes. The real counting result is 211, our method counting result is 207, the Yolov3 result is 170, Yolov4 result is 191, Yolov5s result is 194, Yolov5m result is 200, Faster R-CNN result is 161, and the SSD result is 46. The detection results indicate that Faster-RCNN is not good for images with complex environment backgrounds, Yolov3 and Yolov4 are not good for images with similar background color and occlusion spikes, and the counting effect of SSD is very bad. Additionally, our method is most concentrated in the counting error, mainly between −5 and 10. Therefore, our method is superior to the four methods and comparable to YoLov5.

For the detection results of 60 images, the comparison among "Faster-RCNN," "Yolov3," "Yolov4," "Yolov5s," "Yolov5m," "SSD," and "our method" is performed (**Figure 11**). The RMSE and RMSPE of our method are 1.96 and 0.06. Faster R-CNN is 3.14 and 0.07, Yolov3 is 2.56 and 0.08, Yolov4 is 2.13 and 0.14, Yolov5s is 1.71 and 0.12, Yolov5m is 1.53 and 0.06, and SSD is 10.3 and 0.26. The results indicate that our method has better detection and counting effect than Faster R-CNN, Yolov3, Yolov4, and SSD in the mixed dataset.

CONCLUSION

In this article, we developed a wheat spike detection method based on the SpikeRetinaNet to address the issue of small dense object detection and counting in complex scenes. The method consists of three critical steps: use BiFPN to better integrate multiscale features, network refinement by adding a DSA block, and Soft-NMS was used to solve the occlusion problem. In addition, the WSD images are added to enrich the varieties of the wheat dataset. Based on the methodology, mAP of wheat spikes and counted were outputted, with detection rates of 92.62 and 92.88%, respectively. Therefore, the knowledge generated by this study will greatly aid in the detection and counting of wheat spikes in complex field environments and provide technical reference for agricultural wheat phenotype monitoring and yield prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CW and CY performed conceptualization and supervised the manuscript. CW and JW carried out methodology, provided the software, validated the manuscript, carried out formal analysis, participated in writing, reviewing, and editing, and contributed in visualization. HS and XC investigated the study. ZL and JW provided the resources. CY and JW contributed in data curation. CW and HC involved in writing original draft preparation. CW involved in project administration and contributed in funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

The research was funded by the National Natural Science Foundation of China (key program) (no. U19A2061), the National Natural Science Foundation of China (general program) (nos. 11372155 and 61472161), the Natural Science Foundation of Jilin Province of China (no. 20180101041JC), and the Industrial Technology and Development Project of Development and Reform Commission of Jilin Province (no. 2021C044-8).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.821717/full#supplementary-material

REFERENCES

Alkhudaydi, T., and Zhou, J. (2019). "SpikeletFCN: Counting Spikelets from Infield Wheat Crop Images Using Fully Convolutional Networks," in *International Conference on Artificial Intelligence and Soft Computing*, (Cham: Springer), 3–13.

Bhagat, S., Kokare, M., Haswani, V., Hambarde, P., and Kamble, R. (2021). "WheatNet-Lite: A Novel Light Weight Network for Wheat Head Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Cham: Springer), 1332–1341.

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. 10934. [Preprint].

Bodla, N., Singh, B., Chellappa, R., and Davis, L. S. (2017). "Soft-NMS–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, (Cham: Springer), 5561–5569.

Brooks, J. (2019). *"COCO Annotator."* Available online at: https://github.com/jsbroks/coco-annotator.

Cointault, F., Guerin, D., Guillemin, J. P., and Chopinet, B. (2008). In-field Triticum aestivum ear counting using colour-texture image analysis. *N. Z. J. Crop Horticult. Sci.* 36, 117–130.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., et al. (2020). Global Wheat spike Detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat spike detection methods. *Plant Phenom.* 2020:3521852.

Eversole, K., Feuillet, C., Mayer, K. F., and Rogers, J. (2014). Slicing the wheat genome. *Science* 345, 285–287. doi: 10.1126/science.1257983

Fernandez-Gallego, J. A., Kefauver, S. C., Gutiérrez, N. A., Nieto-Taladriz, M. T., and Araus, J. L. (2018). Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. *Plant Methods* 14, 1–12.

Ferrante, A., Cartelle, J., Savin, R., and Slafer, G. A. (2017). Yield determination, interplay between major components and yield stability in a traditional and a contemporary wheat across a wide range of environments. *Field Crops Res.* 203, 114–127. doi: 10.1016/j.fcr.2016.12.028

Girshick, R. (2015). "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, (New Jersey, NJ: IEEE), 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 580–587.

Hasan, M. M., Chopin, J. P., Laga, H., and Miklavcic, S. J. (2018). Detection and analysis of wheat spikes using convolutional neural networks. *Plant Methods* 14, 1–13.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 770–778.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 7132–7141.

Jin, X., Liu, S., Baret, F., Hemerlé, M., and Comar, A. (2017). Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Rem. Sens. Environ.* 198, 105–114. doi: 10.1016/j.rse.2017.06.007

Jocher, G., Nishimura, K., Mineeva, T., and Vilariño, R. (2020). *YOLOv5*. Available online at: https://github.com/ultralytics/yolov5 (accessed July 10, 2020).

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Informat. Process. Syst.* 25, 84–90. doi: 10.1145/3065386

Kulkarni, A. H., and Patil, A. (2012). Applying image processing technique to detect plant diseases. *Int. J. Modern Engine. Res.* 2, 3661–3664.

LabelImg (2015). *Git code*. Available online at: https://github.com/tzutalin/labelImg (accessed December 25, 2015).

Li, J., Li, C., Fei, S., Ma, C., Chen, W., Ding, F., et al. (2021). Wheat ear recognition based on RetinaNet and transfer learning. *Sensors* 21:4845. doi: 10.3390/s21144845

Li, X., Wang, W., Hu, X., and Yang, J. (2019). "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (New Jersey, NJ: IEEE), 510–519.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 2117–2125.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, (New Jersey, NJ: IEEE), 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). "Ssd: Single shot multibox detector," in *European conference on computer vision*, (Cham: Springer), 21–37.

Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., et al. (2019). Ear density estimation from high resolution RGB imagery using deep learning technique. *Agricult. For. Meteorol.* 264, 225–234. doi: 10.1016/j.agrformet.2018.10.013

Mirnezami, S. V., Young, T., Assefa, T., Prichard, S., Nagasubramanian, K., Sandhu, K., et al. (2020). Automated trichome counting in soybean using advanced image-processing techniques. *Applicat. Plant Sci.* 8:e11375. doi: 10.1002/aps3.11375

Misra, T., Arora, A., Marwaha, S., Jha, R. R., Ray, M., Jain, R., et al. (2021). Web-SpikeSegNet: deep learning framework for recognition and counting of spikes from visual images of wheat plants. *IEEE Access* 9, 76235–76247. doi: 10.1109/access.2021.3080836

Neubeck, A., and Van Gool, L. (2006). "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3, (New Jersey, NJ: IEEE), 850–855.

Pound, M. P., Atkinson, J. A., Wells, D. M., Pridmore, T. P., and French, A. P. (2017). "Deep learning for multi-task plant phenotyping," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (New Jersey, NJ: IEEE), 2055–2063.

Redmon, J., and Farhadi, A. (2017). "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 7263–7271.

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. 02767. [Preprint].

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. 1497. [Preprint].

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, (New Jersey, NJ: IEEE), 618–626.

Slafer, G. A., Savin, R., and Sadras, V. O. (2014). Coarse and fine regulation of wheat yield components in response to genotype and environment. *Field Crops Res.* 157, 71–83. doi: 10.1016/j.fcr.2013.12.004

Sun, S., Li, C., Paterson, A. H., Chee, P. W., and Robertson, J. S. (2019). Image processing algorithms for infield single cotton boll counting and yield prediction. *Comp. Electr. Agricult.* 166:104976. doi: 10.1016/j.compag.2019.104976

Tan, M., Pang, R., and Le, Q. V. (2020). "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 10781–10790.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (New Jersey, NJ: IEEE), 9627–9636.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (New Jersey, NJ: IEEE), 7794–7803.

Wang, Y., Qin, Y., and Cui, J. (2021). Occlusion Robust Wheat Ear Counting Algorithm Based on Deep Learning. *Front. Plant Sci.* 12:645899. doi: 10.3389/fpls.2021.645899.

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, (New Jersey, NJ: IEEE), 3–19. doi: 10.1007/978-3-030-01234-2_1

Yang, Y., Huang, X., Cao, L., Chen, L., and Huang, K. (2019). "Field wheat ears count based on YOLOv3," in *2019 International Conference on Artificial*

Intelligence and Advanced Manufacturing (AIAM), (New Jersey, NJ: IEEE), 444–448.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., et al. (2020). Resnest: Split-attention networks. 8955. [Preprint].

Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., et al. (2021). A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5. *Rem. Sens.* 13:3095. doi: 10.3390/rs13163095

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.