A Method for Wheat Head Detection Based on YOLOv4

Bo Gong (✉ 1410552935@qq.com)

Southwest Minzu University <https://orcid.org/0000-0001-8013-4196>

Daji Ergu

Southwest Minzu University

Ying Cai

Southwest Minzu University

Bo Ma

Southwest Minzu University

Methodology

Keywords: Plant phenotyping, Deep learning, Wheat head detection, CSPNet, GIoU

Posted Date: October 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-86158/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A Method for Wheat Head Detection Based on Yolov4

Bo Gong^{1*}, Daji Ergu¹, Ying Cai¹, Bo Ma¹

*Correspondence: 1410552935@qq.com

1. School of Electronic Information, Southwest Minzu University Chengdu 610225, Sichuan

Abstract

Background: Plant phenotyping by deep learning has increased attention. The detection of wheat head in the field is an important mission for estimating the characteristics of wheat heads such as the density, health, maturity, and presence or absence of awns. Traditional wheat head detection methods have problems such as low efficiency, strong subjectivity, and poor accuracy. However, with the development of deep learning theory and the iteration of computer hardware, the accuracy of object detection method using deep neural networks has been greatly improved. Meanwhile, the detection speed has also been greatly improved. Therefore, using a deep neural network method to detect wheat heads in images has a certain value.

Results: In this paper, a method of wheat head detection based on deep neural network is proposed. Firstly, for improving the backbone network part, two SPP networks are introduced to enhance the ability of feature learning and increase the receptive field of the convolutional network. Secondly, the top-down and bottom-up feature fusion strategies are applied to obtain multi-level features. Finally, we use Yolov3's head structures to predict the bounding box of object. The results show that our proposed detection method for wheat head has higher accuracy and speed. The mean average precision of our method is 94.5%, and the detection speed of our proposed method is 88fps.

Conclusion: The proposed deep neural network can accurately and quickly detect the wheat head in the image which is based on Yolov4. In addition, the training dataset is a wheat head dataset with accurate annotations and rich varieties, which makes the proposed method more robust and has a wide range of application values. The proposed detector is also more suitable for wheat detection task, with the deeper backbone networks. The use of spatial pyramid pooling (SPP) and multi-level features fusion, which all play a crucial role in improving detector performance. Our method provides beneficial help for the breeding of wheat

Keywords: Plant phenotyping, Deep learning; Wheat head detection; CSPNet; GIoU

Background

Wheat is one of the most planted grains in the world, almost all of which is produced for consumption, and is has high nutritional value. The Food and Agriculture Organize

shows that the global wheat production is 756.51 million tons in 2018[1]. These data suggest that the development of the global wheat industry is crucial to global food security and directly affects social stability. Breeding is particularly important to ensure a stable wheat yield. Predicting the yield of wheat is a key step in the breeding work. The traditional counting methods rely on manual observation, which is too subjective and has obvious defects, so that the measurement error is about 10%.

At present, image processing technology and shallow learning are mainly used to detect wheat head. Using the color, texture, shape and other characteristics of the wheat ear itself, an image classifier is constructed to complete automatic detection. For example, Du Shiwei et al.[2] proposed a parabolic segmentation method for wheat ears based on image processing technology. Liu Tao et al.[3] proposed a shallow image segmentation method, which is mainly aimed at wheat ears in a field environment to segment target adhesion areas. Chen Han et al. [4] used the Sobel operator to detect the edges of wheat ears and segmented wheat ears and weeds. Zhao Feng [5] applied the wheat color segmentation method to eliminate the unexpected regions of wheat, and then used the AdaBoost algorithm to classify and locate the wheat ear regions. The above types of detection methods rely on a large number of artificial feature designs and require certain experience. The detection accuracy is easily affected by noises such as light, angle, leaf color, and soil. Therefore, the detection accuracy needs to be improved.

With the development of deep learning theory and the improvement of hardware performance, deep learning has become the most advanced method of computer vision. Now, the popular computer vision tasks, such as object detection, instance segmentation, and semantic segmentation, all use deep neural networks. As the basic task of computer vision, object detection has developed a series of excellent deep learning models, for example, R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], FCOS [9]. The current object detection frameworks are mainly divided into two types: single-stage and two-stage. The representatives of two-stage detectors include R-CNN[6], Fast R-CNN [7], Faster R-CNN [8], etc. R-CNN (Regional Convolutional Network) applies deep learning to the field of object detection, laying a foundation for two-stage target detection. In the selection of region proposals, R-CNN uses the selective search algorithm [14]. In the classification stage, the SVM algorithm is applied. Girshick et al. [7] proposed the Fast R-CNN on the basis of R-CNN, and its innovation was that it eliminated the need to send all candidate boxes into the convolutional neural network. However, this method only needed to send one picture to the network. According to the previous experience, Ren et al.[8] proposed a two-stage object detection method, Faster R-CNN, with faster detection speed. This method introduces the region proposal network (RPN) which extracts candidate bound boxes by setting anchor boxes with different proportions, and realizes an end-to-end network. The representatives of single-stage object detection methods include Yolo series models (Yolo[15], Yolo9000[16], YOLOv3[17], YOLOv4[12]) and SSD. One advantages of the single-stage target detection algorithm is fast detection speed. For example, the detection speed of Yolo can reach 45fps. The idea of Yolo is to

divide the input image into an $S \times S$ grid, and the grid generate a certain number of bounding boxes when the center of an object falls into it. Finally, frame and classify the objects by using NMS (non-maximum suppression) to select the appropriate prediction bounding boxes. According to the previous research of Yolo, Redmon et al. [16] proposed a novel method, Yolo9000. The main innovation of Yolo9000 is the use of multiple computer vision techniques, such as batch normalization, high resolution classifier, location prediction, etc. The detection accuracy of Yolo9000 is 78.6% on VOC2007 dataset. The backbone of Yolo9000 is Darknet-19, with a 3×3 convolutional kernel and global average pooling, which reduces the computational complexity and parameters of the model. Redmon et al. [17] put forward YOLOv3, which used the Darknet-53 network as the backbone network, and introduced the FPN [18] network to achieve the purpose of multi-scale integration.

YOLOv4 [12] was proposed by Alexey et al. The original intention of YOLOv4 was to optimize parallel computing and improve the speed of object detection. The author divided the model into three parts: backbones, neck and heads. The function of each part is different, the backbone part mainly extracts the features. The neck is used to fuse the features extracted from the main part. The role of the head is to predict, including predicting the bounding boxes and the object classification.

The task of wheat head detection can be solved based on deep neural network greatly improve the efficiency of counting, reduce manual participation, and assist in the estimation of wheat yield. And these models have a high generalization ability, which reduce the pre-processing work of image and the dependence on experience. Deep neural networks can well promote the intelligent development of agricultural production. Hasan et al. [10] used the R-CNN network for training, and the average accuracy of wheat ear detection was 93.4%. Zhang Lingxian et al. [11] used a convolutional neural network to construct a winter wheat detection and counting system. Although the use of deep learning technology has obtained

Fig. 1 Wheat head detection

good performance, there are still serious problems. There is always a problem to tradeoff between the detection speed and the detection accuracy. The current methods for wheat head detection still have this disadvantage. In terms of data set, there are also some problems, for example, insufficient datasets, not taking into account the type of wheat, region, growth period, etc.

In this paper, we follow the latest development technology in the field of deep neural network, and propose a novel method for wheat head detection based on the object detection algorithm Yolov4 [12]. We improve the backbone network and adds a spatial pyramid pooling layer (SPP) that can increase the receptive field. Meanwhile, we use a cross-path aggregation network (CSPNet) to intergrade the multi-level features. In this paper, we use the latest global wheat head datasets GWHD [13] to train our proposed method. Our proposed method can detect wheat heads quickly and accurately, and also has good ability of generalization.

Methodology

Overview of Yolov4

The backbone network part of Yolov4 applies the CSPDarknet5 [19]. The CSPDarknet53 network is a CSPNet (Cross Stage Partial Network) network added on the basis of Darknet53 [17]. Darknet53 draws on the idea of ResNet [20], namely residual connection, to ensure that the network has depth while also alleviate the vanishing gradient problem. CSPNet can enhance the learning ability of CNN, while reducing the amount of calculation and memory cost. A good detector should have a larger receptive field. The neck of the Yolov4 network uses two networks, SPP [21] and PANet [22]. The SPP network applied in the neck can effectively increase the receptive field and help separate contextual features. The PANet (Path Aggregation Network) plays a role in shortening the path connecting low-level information and high-level information, and converging parameters at different levels. The Yolov4 network head inherits the head structure of Yolov3.

The head predicts the bounding box of the object, and outputs the center coordinates, width, and height, i.e., $\{x_{center}, y_{center}, w, h\}$. Then the expression of the predicted bounding box is shown as follows:

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w \cdot e^{t_w} \\ b_h &= p_h \cdot e^{t_h} \end{aligned} \tag{1}$$

where p_w and p_h represent the width and height of the prior bounding box, respectively. (c_x, c_y) is the coordinate of the top left corner of the image. Figure 3 shows the size of the prior bounding box and the position of the predicted bounding box.

Fig. 2. Spatial pyramid network (SPP)

Fig. 3 Conversion of true bounding box and predicted bounding box.

Fig. 4 The structure of YOLOv4.

Our proposed wheat head detector

In this paper, our detector is mainly composed of three parts: backbones, neck and heads. For the backbone, in order to improve the receptive field of the network, spatial

pyramid pooling [21] (SPP, Spatial Pyramid Pooling) is applied. Owing to the big size of image in this article is 1024×1024 , using scaling and cropping operations will make the picture with more noise. To solve this problem, we add a spatial pyramid pooling network to the front of the backbone network. In addition, a fixed-size feature vector and the original image information are output effectively. The spatial pyramid pooling network is shown in Figure 2, which mainly solves the multi-size problem of the input image through multi-scale pooling.

An excellent backbone network should learn as many features of pictures as possible. In order to enhance the learning ability of the backbone network, A Cross Stage Partial Network (CSPNet) [19] is applied. CSPNet was proposed by Wang et al. to enhance the learning ability of convolutional neural networks. This network can still maintain or enhance the learning ability of the CNN network while reduce the amount of calculation by 20%. Therefore, in order to enhance the learning ability of the backbone network, we use the CSPNet network and obtains a new network structure, CSPDarkNet53 [19]. The structure is shown in Figure 5 and consists of two parts: skip connection and main part. The main part retains the original DarkNet53 structure, and multiple residual blocks are stacked. The skip connection part is directly connected to a concat layer of the network, and it is also spliced with the main part.

In this paper, SPP [21] is introduced to the tail of backbone network, denoted as SPP-2. In addition, the purpose of introducing SPP-2 is different from SPP-1. This part is to separate the context features and facilitate the neck network fusing the global feature information. Firstly, the SPP-2 network performs a convolution operation on the input features of the upper layer. And then, it performs a maximum pooling operation of different scales. The pooling size of Pool-1, pool-2, and pool-3 are 5, 7, and 13 respectively. In addition, the step size is 4, 6, and 13 respectively. SPP-2 merges the output features of the three pooling layers and inputs them to the next convolutional module to perform feature learning to obtain the abundant local features.

Fig. 5 The structure of CSPDarkNet53 network.

Path Aggregation Network (PANet) is added to integrate multi-level features. The low-level information contains the outline of object. On the contrary, the high-level features contain the details of target. In the field of object detection, the part of detector which collects feature maps is usually called the neck. The neck usually consists of a bottom-up path and a top-down path. In view of the particularity of the task in this paper,

the photos of wheat heads contain obvious structural features and abundant details features. We use PAN (Path Aggregation Network) as the neck of the detector to collect multi-level features and connect with the spatial pyramid network (SPP-2) to form a bottom-up and top-down combination. The path aggregation network is mainly composed of a series of convolution, pooling, upsampling, and splicing layers. The input of the Path Aggregation Network comes from three parts, two of which are from the feature layer of the backbone network, and the other input is from the feature layer of SPP-2. The output of the path aggregation network is used as the input for the head of object detection network.

The model in this paper uses Yolov3 as the head to predict the bounding box. First, calculate the coordinates, width, and height of the prediction boxes according to formula 1. Secondly, the confidence threshold is set to filter out the prediction frames with low scores. Finally, non-maximum value suppression is used to determine the final prediction frame. Figure 3 also shows the principle of calculating the position of the bounding box.

Loss Function

The current object detection method uses IOU (Intersection over Union) to determine the degree of overlap between the predicted and the ground-truth bounding box. IOU is represented as follow:

$$IOU = \frac{M \cap N}{M \cup N} \quad (2)$$

where M is the prediction bounding box, represented by $(x_{center}, y_{center}, w, h)$. N is the ground-truth bounding box (x, y, w, h) . However, this optimization method has the disadvantage of not being able to optimize non-overlapping parts. Therefore, we introduce the generalized GIOU [23] loss function, represented as follow:

$$GIOU = IOU - \frac{|A_c - U|}{|A_c|} \quad (3)$$

where A_c represents the minimum bounding box between the predicted bounding box and the ground-truth bounding box. U is the union of the predicted and the ground-truth bounding boxes, i.e., $M \cup N$. The loss function not only pays attention to the overlapping area, but also focuses on the non-overlapping area of the two kinds of boxes which better reflects the overlap of the two boxes. As shown in Figure 7, the intuitive schematic diagrams of IOU and GIOU, respectively. The bounding box regression loss function used in this article is:

$$Loss_{GIOU} = 1 - GIOU \quad (4)$$

The value range of $GIOU$ is $(-1, 1)$. The higher the overlap of the bounding box M and N , the closer the $GIOU$ is to 1. When M and N do not overlap, optimization can still be performed, which benefits from the existence of the smallest bounding boxes, i.e., A_c . By contrast, this advantage is missing from IOU.

Fig. 6 The structure of our proposed method in this paper

Fig. 7 GIOU and IOU

Experimental and Results

Dataset

The datasets applied in this paper is the Global Wheat Head Detection dataset GWHD [13]. The GWHD dataset was constructed collaboratively by numerous countries. The GWHD dataset is the first large-scale dataset to detect wheat heads from field optical images. The wheat head pictures are varieties grown in different regions. The dataset uses the web-based annotation tool coco annotator [24]. The platform is rich in features, with all the tools required to label objects. Labeling the large density bounding boxes is difficult. Therefore, it is required to draw a box containing all the pixels of the wheat head when the image is complete or the part of the veins is occluded.

The labelled part contains at least one wheatear. Figure 8 shows the ground-truth boxes and label files of part of the dataset. The label information is the top-left coordinates (x_{min}, y_{max}) of the bounding box and the width and height of the bounding box w, h .

(a)The image of dataset with ground-truth bounding boxes

```
[627.0, 302.0, 122.0, 75.0]
[412.0, 367.0, 68.0, 82.0]
[953.0, 220.0, 56.0, 103.0]
[30.0, 70.0, 126.0, 133.0]
[35.0, 541.0, 46.0, 46.0]
[103.0, 60.0, 117.0, 83.0]
[417.0, 4.0, 110.0, 91.0]
[764.0, 299.0, 119.0, 93.0]
[539.0, 58.0, 58.0, 130.0]
[139.0, 274.0, 121.0, 76.0]
[461.0, 634.0, 118.0, 64.0]
```

(b)The annotated file for the dataset (.txt)

Figure 8. An example of dataset annotation

Evaluation Metrics

In order to evaluate the effect of wheat head detection, we use the following evaluation metrics:

(1) Recall and precision:

$$P_{recision} = \frac{TP}{TP+FP} \quad (5)$$

$$R_{ecal} = \frac{TP}{TP+FN} \quad (6)$$

True positive (TP) represents that samples are predicted to be correct and actually positive. False positive (FP) represents that samples are predicted to be positive but actually negative. In addition, false negative (FN) represents that samples are predicted to be negative and actually negative.

(2) The AP is to average the precision, and it is shown in equation 4:

$$AP = \int_0^1 p(r)dr \quad (7)$$

In practice, the PR curve is smoothed, and the area under the curve is used to calculate the object's AP value.

(3) Mean Average Accuracy (mAP) means that the average accuracy of all categories is added, and it is divided by the number of categories, shown as follows:

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \quad (8)$$

where N is the number of object classifications. We use two ranges of average precision mean, $map@0.5$, and $map@0.5:0.95$. The $mAP@0.5$ represents the average accuracy of the confidence threshold of 50%, which is recorded as mAP50 in this article. The $mAP@0.5:0.95$ represents the mean value of the average

accuracy in the range of 5%-95% confidence level, which is expressed directly in this article.

(4) Frame Per Second (FPS), means that the number of images can be detected per second, used to evaluate the speed of detector. Only the detection speed of method is fast enough to realize real-time detection and meet the needs of the industry.

Training

This experiment, we use Intel(R) Xeon(R)Silver 4110 CPU, and GPU is GeForce RTX 2080Ti for accelerating model training. The programming language is python3.7 in this paper, based on pytorch1.5.0 deep learning framework. The specific process of this experiment is shown in Figure 9. Main steps of this experiment are as follows: Firstly, we remove some pictures according to the size of the bounding box from the original dataset, keeping the dataset with accurate and clean labels. Then, a series of data enhancement operations are performed, including rotation, cropping, adding noise, etc. Finally, the training datasets are input into the deep neural network for training.

In the training phase of this paper, all models are trained for 150 epochs, using SGD (stochastic gradient descent optimization), and momentum and decay weights are set as 0.937 and 0.0005, respectively. The batch size is 16, and the initial learning rate is 0.01. It can be seen from Figure 10 that the loss function value of the later training is as low as about 0.03. The training accuracy and recall rate change curve are shown in Figure 11.

Fig.9 The work-flow of our proposed detector for wheat head

Fig. 10 GIOU loss function of training

Fig. 11 The curves of precision and recall about training

Table 1 The comparisons of numerous detection methods.

Method	Datasets	mAP ₅₀	mAP ₉₅	FPS
Yolov3	GWHD	90.5%	46.7%	50
Yolov4	GWHD	91.4%	51.2%	37
Faster R-CNN	GWHD	76.6%	49.1%	10
Our method	GWHD	94.5%	54.5%	83

Table 2 Using different backbones for wheat head detector training

Backbones	Datasets	mAP ₅₀	mAP ₉₅	FPS
DarkNet-53	GWHD	87.6%	48.6%	55
CSPDarkNet-53	GWHD	90.2%	50.4%	65
Our backbone	GWHD	94.5%	54.5%	83

Results

It can be seen from Figure 12 that our proposed method performs well for wheat head detection. The detection effect of the wheat head at the mature stage is better, which benefits from the characteristic integrity and uniqueness of wheat head at the mature stage. In order to scientifically show the detection performance of our method, we train and test our proposed method with other detector on the same dataset, and compare the test results. Table 1 shows the detection performance of each detector.

It can be seen from Table 1 that our proposed method in this paper has achieved good results on the task of wheat head detection. Compared with the Yolov3, the *mAP50* and *mAP95* indicators of our method are improved 4% and 7.8%, respectively. In addition, the detection speed also increases 33fps. Compared to Yolov4, our proposed method increases the *mAP50* and *mAP95* up to 5.2% and 3.3%, respectively. the detection speed has increased by 30fps. Compared with the two-stage detection method, Faster R-CNN, our method in this paper does not lose advantage in accuracy. Its *mAP50* and *mAP95* increases by 17.9% and 5.4%, respectively. The speed of our detector increases by 70fps. In general, our proposed method in this paper achieves a good performance in detecting wheat head, and its performance is slightly better than Yolov4. Our method guarantees the speed and accuracy of detector all has a good performance.

In order to illustrate the influence of different backbones on our detector, we experimentize and confirm the three backbone networks. While keeping the neck and head of our method unchanged, DarkNet-53, CSPDarkNet-53, and the backbone of our proposed are used for experiment. It can be seen from Table 2 that compared with the other two backbone networks, the improved backbone in this paper increases *mAP50* by 4.3% and 6.9%, respectively. In addition, our proposed backbone improves *mAP95* by 4.1% and 5.9%, respectively. In terms of detection speed, using the improved backbone is faster than the detector based on DarkNet-53 and CSPDarkNet-53 by 18fps and 28fps, respectively. In conclusion, the improved backbone in this article improves the accuracy significantly, while also achieve better real-time performance. Therefore, our proposed method has applying and researching values in wheat head detection direction.

Fig. 12 Visual results our proposed method on GWHD

Conclusion

In this paper, in order to solve the problem of wheat head detection, we propose an improved object detection method. First, in order to increase the receptive field of backbone network and enhance the ability of feature learning of the backbone network, the advantages of the SPP network and the CSPNet are absorbed. Therefore, our proposed backbone network is constructed with including SPP and CSPDarkNet-53. Secondly, a spatial pyramid module is added at the bottom of backbone mainly to obtain

multi-scale local features. And then, in the terms of neck network, we use a bottom-up to top-down connection strategies to integrate multi-level features. Finally, we use GIOU as a loss function. The experimental results show that our proposed method effectively solves the task of wheat head detection, a multi-variety and multi-season field of wheat images. The mean Average Precision is 94.5%, and the detection speed on GTX2080Ti is 88 FPS. In future work, we will consider updating the method of this paper, and introduce self-attention mechanism to enhance the feature learning ability of detector. Regarding the datasets, we consider using generative adversarial network for data augmentation.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Consent and approval for publication from all the authors was obtained.

Availability of data and materials

The datasets used is publicly available at <http://www.global-wheat.com/>

Competing interests

The authors declare that they have no competing interests.

Funding

Graduate Innovation Project of Southwest Minzu University (No. CX2019SZ16)

Author detail

¹ School of Electronic Information, Southwest Minzu University Chengdu 610225, Sichuan

Authors' contributions

BM was responsible for experimental design. BG was responsible for implementing and executing the deep learning analysis. BG prepared the figures and wrote the first draft of the manuscript. DE and CY revised the manuscript. All authors contributed equally to the final editing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank SMU for the foundation to undertake this research work.

References

1. FAOSTAT. <http://www.fao.org/faostat/en/#home>
2. Shiwei Du, Yinian Li, Min YAO, Ling LI, Qishuo D, Ruiyin HE. Counting method

- of grain number based on wheatear spikelet image segmentation. 2018;
3. Liu Tao, Sun Chengming, Wang Lijian, et al. Journal of Agricultural Mechanization Research. 2013; 21:90-3.
 4. Chen Han, Lv Xingjun, Tian Fengzhen, et al. Wheat Panicle Image Segmentation Based on Sobel Operator-edge Detection. Journal of Agricultural Mechanization Research. 2013; 3:3–6.
 5. Zhao Feng, Wang Kejian, Yuan Yingchun. Study on Wheat Spike Identification Based on Color Features and AdaBoost Algorithm. Crops, 2014;1:1–5.
 6. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014;580–7.
 7. Girshick R. Fast R-CNN. Proc IEEE Int Conf Comput Vis. 2015;2015 Inter:1440–8.
 8. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017; 39:1137–49.
 9. Tian Z, Shen C, Chen H, He T. FCOS: Fully Convolutional One-Stage Object Detection. 2019; Available from: <http://arxiv.org/abs/1904.01355>
 10. Hasan MM, Chopin JP, Laga H, Miklavcic SJ. Detection and analysis of wheat spikes using Convolutional Neural Networks. Plant Methods [Internet]. BioMed Central; 2018; 14:1–13. Available from: <https://doi.org/10.1186/s13007-018-0366-8>
 11. Zhang L, Chen Y, Li Y, Ma J, Du K. Detection and Counting System for Winter Wheat Ears Based on Convolutional Neural Network. Nongye Jixie Xuebao/Transactions Chinese Soc Agric Mach. 2019; 50:144–50.
 12. Bochkovskiy A, Wang C-Y, Liao H-YM. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020; Available from: <http://arxiv.org/abs/2004.10934>
 13. David E, Madec S, Sadeghi-Tehran P, Aasen H, Zheng B, Liu S, 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. 2020;1–16. Available from: <http://arxiv.org/abs/2005.02162>
 14. Uijlings JRR, Van De Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int J Comput Vis. 2013; 104:154–71.
 15. Redmon JSDRGAF. (YOLO) You Only Look Once. Cvpr. 2016;
 16. Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:6517–25.
 17. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. 2018; Available from: <http://arxiv.org/abs/1804.02767>
 18. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:936–44.
 19. Wang C-Y, Liao H-YM, Yeh I-H, Wu Y-H, Chen P-Y, Hsieh J-W. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. 2019; Available from: <http://arxiv.org/abs/1911.11929>

-
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.* 2016;2016-Decem:770–8.
 21. He K, Zhang X, Ren S, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. :1–14.
 22. Liu S, Qi L, Qin H, Shi J, Jia J. Path Aggregation Network for Instance Segmentation. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.* 2018;8759–68.
 23. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.* IEEE; 2019;2019-June:658–66.
 24. Brooks J. COCO Annotator. 2019.

Figures

Figure 1

Wheat head detection

Figure 2

Spatial pyramid network (SPP)

Figure 3

Conversion of true bounding box and predicted bounding box.

Figure 4

The structure of YOLOv4.

Figure 5

The structure of CSPDarkNet53 network.

Figure 6

The structure of our proposed method in this paper

Figure 7

GIoU and IOU

(a)The image of dataset with ground-truth bounding boxes

```
[627.0, 302.0, 122.0, 75.0]
[412.0, 367.0, 68.0, 82.0]
[953.0, 220.0, 56.0, 103.0]
[30.0, 70.0, 126.0, 133.0]
[35.0, 541.0, 46.0, 46.0]
[103.0, 60.0, 117.0, 83.0]
[417.0, 4.0, 110.0, 91.0]
[764.0, 299.0, 119.0, 93.0]
[539.0, 58.0, 58.0, 130.0]
[139.0, 274.0, 121.0, 76.0]
[461.0, 634.0, 118.0, 64.0]
```

(b)The annotated file for the dataset (.txt)

Figure 8

An example of dataset annotation

Figure 9

The work-flow of our proposed detector for wheat head

Figure 10

GIOU loss function of training

Figure 11

The curves of precision and recall about training

Figure 12

Visual results our proposed method on GWHD