Plant Methods

METHODOLOGY

Open Access

# Improving multi-scale detection layers in the deep learning network for wheat spike detection based on interpretive analysis

Jiawei Yan[1,2†], Jianqing Zhao[1,2†], Yucheng Cai[1,2], Suwan Wang[1,2], Xiaolei Qiu[1,2], Xia Yao[1,2,3], Yongchao Tian[1,4], Yan Zhu[1,2], Weixing Cao[1,2] and Xiaohu Zhang[1,2,4*]

## Abstract

**Background** Detecting and counting wheat spikes is essential for predicting and measuring wheat yield. However, current wheat spike detection researches often directly apply the new network structure. There are few studies that can combine the prior knowledge of wheat spike size characteristics to design a suitable wheat spike detection model. It remains unclear whether the complex detection layers of the network play their intended role.

**Results** This study proposes an interpretive analysis method for quantitatively evaluating the role of three-scale detection layers in a deep learning-based wheat spike detection model. The attention scores in each detection layer of the YOLOv5 network are calculated using the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm, which compares the prior labeled wheat spike bounding boxes with the attention areas of the network. By refining the multi-scale detection layers using the attention scores, a better wheat spike detection network is obtained. The experiments on the Global Wheat Head Detection (GWHD) dataset show that the large-scale detection layer performs poorly, while the medium-scale detection layer performs best among the three-scale detection layers. Consequently, the large-scale detection layer is removed, a micro-scale detection layer is added, and the feature extraction ability in the medium-scale detection layer is enhanced. The refined model increases the detection accuracy and reduces the network complexity by decreasing the network parameters.

**Conclusion** The proposed interpretive analysis method to evaluate the contribution of different detection layers in the wheat spike detection network and provide a correct network improvement scheme. The findings of this study will offer a useful reference for future applications of deep network refinement in this field.

**Keywords** Wheat spike detection, Deep learning network, Attention score, Interpretive analysis

† Jiawei Yan and Jianqing Zhao have contributed equally to this work and should be regarded as co-first authors.

*Correspondence:
Xiaohu Zhang
zhangxiaohu@njau.edu.cn
[1] National Engineering and Technology Center for Information Agriculture, Nanjing Agricultural University, Nanjing 210095, China
[2] Key Laboratory for Crop System Analysis and Decision Making, Ministry of Agriculture and Rural Affairs, Nanjing 210095, China
[3] Jiangsu Key Laboratory for Information Agriculture, Nanjing 210095, China
[4] Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing 210095, China

## Introduction

Wheat is one of the world's important food crops. The statistics of the Food and Agriculture Organization of the United Nations show that global wheat production in 2021 is 776.8 million tons with a planted area of 220 million hectares, and the global wheat production in 2022 is expected to be 770.8 million tons [1]. In the context of world population growth and global climate change, ensuring stable and increased wheat production is crucial to world food security. Meanwhile, because the number of wheat spikes per acre and grain weight per spike directly determine the final yield [2], detecting and counting wheat spikes are important [3] for predicting and measuring wheat yield before harvest.

With the improvement of computer technology in recent years, deep learning-based object detection techniques have been increasingly applied to wheat spike detection. Some are two-stage detection methods, e.g., R-CNN [4], Fast-RCNN [5], and Faster-RCNN [6]. Some are one-stage detection, e.g., YOLO (You only look once) [7], YOLO9000 [8], YOLOv3 [9], YOLOv4 [10], and YOLOv5 [11].

Based on these technical means, some researchers evaluated the existing methods on public datasets [12, 13], while others focused on improving the state-of-art deep-learning-based methods on their private datasets [14, 15]. In these datasets, all wheat spikes have corresponding ground-truth boxes.

Both non-convolutional and convolutional wheat spike detection methods focus on wheat spike size information. Some non-convolutional methods use image processing technology and machine learning to design feature extraction for small-sized wheat spikes detection [16, 17]. Due to the differences in variety, environment, and observation scenarios, the size of wheat spikes in images varies significantly, resulting in different roles of multi-scale detection layers of the neural network in wheat spike detection. The problem is how to quantitatively analyze the role of multi-scale detection layers in the complex network structure. Solving the problem will provide a correct direction for optimizing the wheat spike detection network, and it can also provide a reference for the research of multi-size object detection [18, 19]. The development of frontier deep learning interpretive techniques provides a reliable technical way to study this kind of problem. Selvaraju et al. proposed a method based on Gradient-weighted Class Activation Mapping (Grad-CAM), which can use the gradient information in the network back propagation along with network feature layers to generate a "visual interpretation" as the reason for decision-making of the deep learning model. It can generate the attention areas of the network layer to the specific object and use heat maps with location and semantic information to highlight the important areas in the image for predicting the conceptual object. This method provides users with explanatory results and helps them to successfully identify or optimize the stronger deep learning network [20]. In this study, we introduce a network attention method based on the Grad-CAM algorithm to explore the role of multi-scale detection layers in the deep learning model and refine the wheat detection network model according to the interpretive analysis results. First, we trained a wheat spike detection model based on YOLOv5. Then, the attention scores of each network detection layer were quantified based on the Grad-CAM algorithm. Following that, We obtained the performance of different detection layers for the detection of wheat spikes and finally clarified the optimal improvement direction of the network. In addition, the optimized wheat spike detection network was successfully constructed and validated on the Global Wheat Head Detection (GWHD) dataset.

## Methods

### Overall technical framework

This study proposes a strategy for improving the detection layer scales of a deep learning-based wheat spike detection network based on interpretive analysis (Fig. 1). YOLOv5 is applied as the basic wheat spike detection network [11]. First, the feature maps of all channels in each detection layer are obtained, and the backpropagation of the detection network is performed to obtain the gradient values of each feature map to calculate the weights of feature maps. Weighted summation is conducted between the weight values and the feature map values. This process involves calculating the mean value of all pixels in an individual feature channel at each detection layer. This mean value is used as the weight parameter for the corresponding feature map. The calculated weight parameter is multiplied by the pixel values of the corresponding feature map. The results of these calculations are summed across all feature channels in the detection layer and weights of feature maps can be obtained. Second, the Grad-CAM value is input to the ReLU activation function to obtain the positive class activation mapping. Thus, the wheat spikes attention area of the network is obtained. Then, the attention score of each detection layer for an individual wheat spike is quantified by comparing the prior wheat spike labeled box and its attention area. Finally, the attention score in each detection layer is assessed to improve detection layer scales. With this improvement, a stronger wheat spike detection network is constructed to achieve higher performance in wheat spike detection.

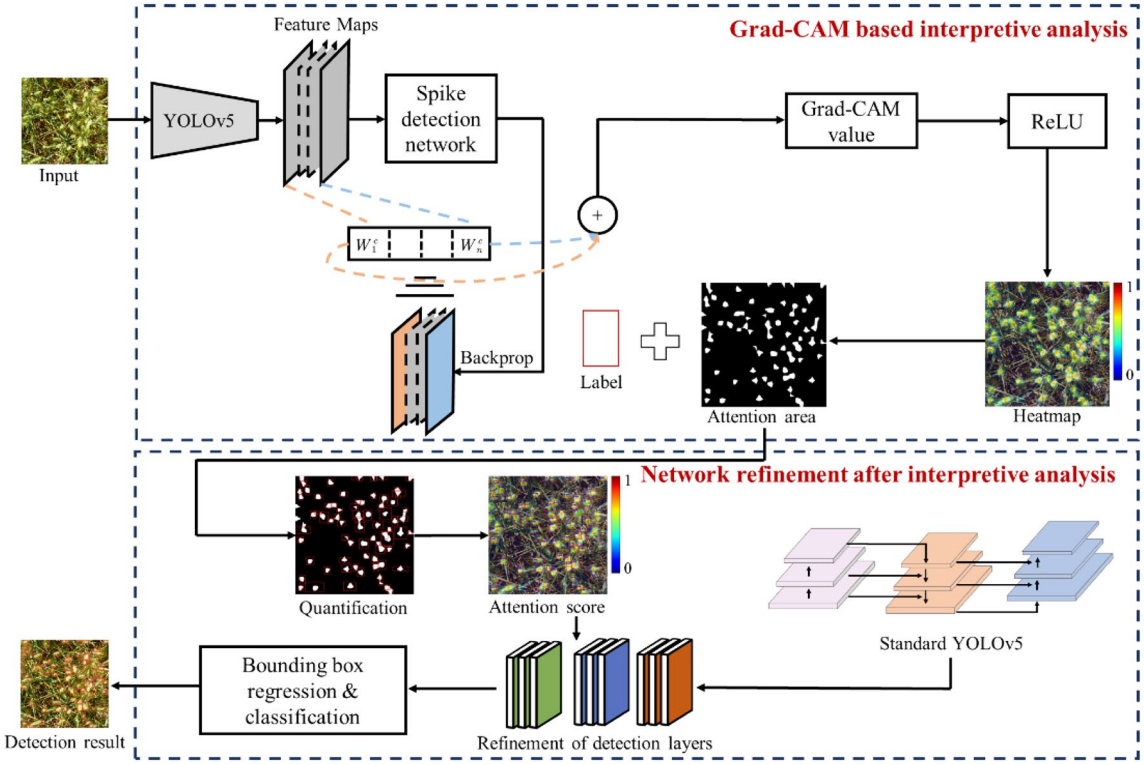


**Fig. 1** Technical framework

### Global wheat head detection (GWHD) dataset

GWHD dataset is an important large wheat spike image dataset in the world, covering a total of 6515 images of wheat spikes from 12 different countries, with different growth stages, genotypes, planting conditions, and image acquisition methods. The spatial resolution of images is 1024×1024, and the spectral bands are red, green, and blue. The total number of manually labeled wheat spikes in the dataset is 275187. The dataset consists of two versions, GWHD_2020 [21] and GWHD_2021 [22]. Among them, GWHD_2021 is an adaptation and expansion of GWHD_2020. In this study, GWHD_2021 and GWHD_2020 were both used in model development. In particular, based on the wheat spikes size distribution in the GWHD dataset, 1000 representative wheat spike images from four sub-datasets (ethz_1, arvalis_1, usask_1, and inrae_1) in GWHD_2020 were evenly selected for analyzing the role of multi-scale detection layers in the network, totaling 44538 wheat spikes. The differences in wheat spike morphology and size in the four sub-datasets are significant (Table 1, Fig. 2).

### Improvement of detection layer scale in standard YOLOv5

#### *Overview of YOLOv5*

The study adopts the YOLOv5 object detection model as the benchmark network. YOLOv5 is a high-performance one-stage deep learning framework. It consists of four main modules, including the input module, the backbone

**Table 1** The selected images from GWHD_2020 for interpretive analysis

| Sub-dataset | Average spike width (Pixels) | Average spike length (Pixels) | Average spike size (Pixels) | Number of selected images | Number of labeled spikes |
|---|---|---|---|---|---|
| arvalis_1 | 80 | 75 | 6312 | 300 | 12,800 |
| inrae_1 | 120 | 119 | 15271 | 176 | 3701 |
| usask_1 | 98 | 87 | 9247 | 200 | 5737 |
| ethz_1 | 76 | 63 | 4783 | 324 | 22,300 |
| Total | – | – | – | 1000 | 44,538 |


**Fig. 2** Sample wheat spike images of the four sub-datasets of GWHD_2020

module, the neck module, and the detection module. The backbone module of YOLOv5 is mainly responsible for the feature extraction of wheat spikes. The neck module of YOLOv5 focuses more on image feature extraction and fusion than the backbone module. With Path Aggregation Network (PANet) and Bi-directional Feature Pyramid Network (BiFPN), the neck module achieves bottom-up and top-down feature fusion by two up-sampling operations [23, 24]. The detection module of YOLOv5 conducts object bounding box generation and object prediction in three scales: small-scale, medium-scale, and large-scale. Correspondingly, the standard YOLOv5 network contains three essential network layers: small-scale detection layer, medium-scale detection layer, and large-scale detection layer. In this study, to improve the network, we use interpretive analysis of these different scale detection layers on the performance of the wheat spike detection.

#### *Interpretive analysis of detection layer scale based on Grad-CAM*

We use the Grad-CAM algorithm to extract the attention areas of wheat spikes from the pre-trained wheat spike detection network on three-scale detection layers. Then, the quantitative attention scores of all wheat spikes can be obtained by comparing the prior wheat spike labeled boxes with the attention areas of wheat spikes. Finally, the contribution of each detection layer of the network to the successful detection can be quantitatively evaluated based on the attention score after calculating the proportion of wheat spikes in each score interval. Grad-CAM is a visualization interpretation method for neural networks [25]. The principle of Grad-CAM is similar to the other class activation mapping (CAM) methods. It calculates $\alpha_k$ the average value of the gradients in each channel $k$ of the network feature layer as weights [26]:

$$a_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y}{\partial A_{ij}^k} \tag{1}$$

Where $y$ is the prediction score of the network for the wheat spike class; $A_{ij}^k$ represents the value of the *i-th* row and *j-th* column in the feature map of channel $k$; $Z$ represents the multiplied value of width and height of the feature map.

Then, the weighted summation operation of weight $\alpha_k$ and feature map $A_k$ is performed on these channels, and the ReLU activation function filters out the negative values of the feature layer to obtain the final Grad-CAM value $L_{Grad-CAM}$:

$$L_{Grad-CAM} = \mathrm{ReLU}\left(\sum_k \alpha_k A^k\right) \tag{2}$$

where $A_k$ represents the *k-th* channel in the feature layer $A$; $\alpha_k$ represents the weight of the *k-th* channel in the feature layer.

Grad-CAM values are visualized in heat maps, thus visualizing the role of important areas in the network on wheat spike detection. Meanwhile, the Grad-CAM value is applied to extract the attention area to wheat spikes of three-scale detection layers. We define the attention area $R_{Grad-CAM}$ as the region with a non-empty Grad-CAM value area and is quantitatively compared with the area $R_{Label}$ f prior wheat spike labeled boxes. Then the $S$ value is derived as the contribution of detection layers to the wheat spike detection:

$$S = \frac{R_{Grad-CAM}}{R_{Label}} \tag{3}$$

where $R_{Grad-CAM}$ represents the area of attention region; $R_{Label}$ epresents the area of wheat spike labeled boxes; $S$ represents the final attention score.

#### *Network improvements*

The proposed method quantitatively evaluates the performance and contribution of the original three detection layers based on the attention scores. Based on this evaluation, we develop a network improvement strategy. Removal of the large-scale detection layer will be

considered when its attention score is poor. Adding a micro-scale detection layer will be considered to improve the detection of small-sized objects when the attention score of the small-scale detection layer is excellent [27]; otherwise, it is removed. When the attention score of the medium-scale detection layer is excellent, the feature enhancement operation will be applied. In the Yolov5 backbone, shallow convolutional layers can extract spatial features, while deep convolutional layers can extract semantic features. The semantic and spatial features are upsampled and downsampled, respectively, and combined bidirectionally by fusion. The multi-scale features are then directed to the medium-scale detection layer. This process introduces more feature information with the same scale from the backbone module to the neck module and enhances the features in the medium-scale detection layer. These measures build a new strong detection network for wheat spike objects (Fig. 3).

## Experimental settings

### *Multi-resolution training strategy*

The study adopts a multi-resolution training strategy. The network is trained by inputting images with different resolutions of 150×150, 300×300, 450×450 and 600×600 to obtain the trained model.

The experiment is conducted on a workstation equipped with Intel® i7 10,700 processor, NVIDIA® Geforce GTX 1080Ti graphics processor (12GB memory), 32GB RAM, and 1TB storage. The computer operating system is Ubuntu 16.06, and the hyperparameter settings for network training are set as follows (Table 2). Batch size, training round, learning rate, and momentum are separately set to 8, 100, 0.01, and 0.9.

### *Evaluation metrics*

The study adopts *precision*, *recall,* and average precision (*AP*) to evaluate the performance of the deep-learning network model for wheat spike detection. The *precision* and *recall* are defined as:

$$precision = \frac{TP}{FP + TP} \tag{4}$$

$$recall = \frac{TP}{FN + TP} \tag{5}$$

**Table 2** Network training hyperparameter setting

| Input size | Batch size | Epoch | Learning rate | Momentum |
|---|---|---|---|---|
| 150×150 | 8 | 100 | 0.01 | 0.9 |
| 300×300 | 8 | 100 | 0.01 | 0.9 |
| 450×450 | 8 | 100 | 0.01 | 0.9 |
| 600×600 | 8 | 100 | 0.01 | 0.9 |

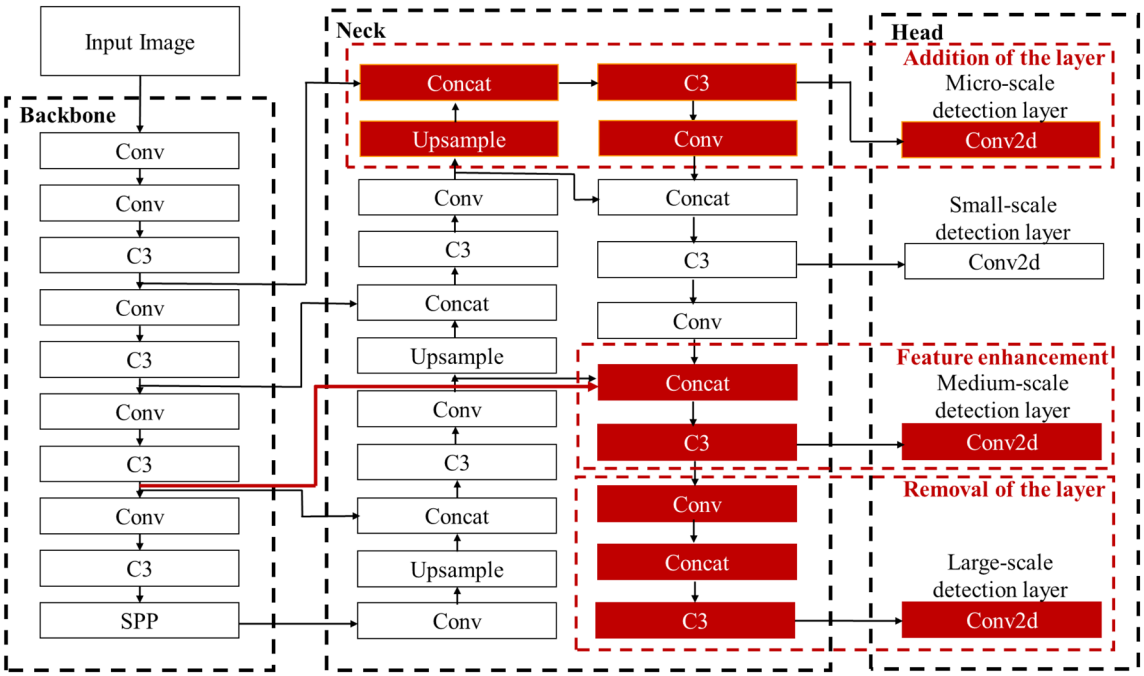


**Fig. 3** Measures of improving the detection layer scale for wheat spike detection: white parts represent standard YOLOv5, and red parts represent network improvements

where *TP*, true positive, means that positive samples are correctly predicted as positive; *FP*, false positive, means that negative samples are incorrectly predicted as positive; *FN*, false negative, means that positive samples are incorrectly predicted as negative.

Since *precision* and *recall* are a pair of indicators that affect each other, it is difficult to fully evaluate the network using one of the two indicators alone. Therefore, the average precision (*AP*) is introduced. *AP* is the average *precision* of *recall* in the 0–1 interval for detecting a certain class of objects and obtained by:

$$AP = \int_0^1 precision(recall) drecall \tag{6}$$

## Results

### Attention scores of each detection layer

Experimental results show that attention areas of wheat spikes in the small-scale detection layer were small, and therefore the calculated attention scores are relatively low (Fig. 4a, d).

Moreover, the statistics of attention scores of the small-scale detection layer show a trend of higher scores for smaller sizes and weaker attention for larger wheat spikes (Table 3, Fig. 5). Most wheat spikes have low attention scores in the range of 0.0–0.1 and 0.1–0.2 (64.4% and 27.5%). They have 8682 and 3438 pixels in size. The proportion of wheat spikes in the 0.2–0.7 score interval is small, while these wheat spikes are also small, with a size below 2137 pixels. Moreover, there are no wheat spikes in the score interval 0.8–1.0.

In Fig. 4, 75 wheat spikes are labeled, and attention areas of wheat spikes are visually larger and more accurate in the medium-scale detection layer than in the small-scale detection layer (Fig. 4b, e). Attention scores of the medium-scale detection layer are more evenly distributed and show a trend of higher scores for smaller wheat spikes (Table 3, Fig. 5). Moreover, most wheat spikes have a moderate score in the range of 0.1 to 0.4. These wheat spikes also have a medium size. The proportion of wheat spikes in the 0.4–1.0 interval is small, and the sizes of wheat spikes in this interval are smaller, below 3360 pixels.

Attention scores of the large-scale detection layer indicate that this layer weakens wheat spike detection (Fig. 4c, f). The largest proportion of wheat spikes is in score interval 0.0–0.1, with a proportion of 38.8%. Besides, these spikes are small, with an average size of 3891 pixels. In the remaining score intervals, wheat spikes are evenly distributed (Table 3, Fig. 5). Attention areas of wheat spikes in the large-scale detection layer are visually large. There are many wheat spike labeled boxes without existing attention areas. In addition, there is a phenomenon that attention areas exceeded labeled boxes. It indicates that the network confuses background areas with wheat spike areas. Therefore, the network cannot make accurate inferences (Fig. 6).

Wheat spikes with an attention score of 0 in three detection layers are counted (Table 4). The three detection layers have 17.5%, 4.0%, and 30.0% wheat spikes with an attention score of 0. It indicates that the deep network failed to focus on this part of areas where existing wheat spikes. Three parts of wheat spikes have average sizes of 13040, 4660, and 3434 pixels. The small-scale detection layer has difficulty identifying larger wheat spikes, while the large-scale detection layer has difficulty identifying smaller ones. Medium-scale detection layer merely ignores 4% of all wheat spikes. It achieves the best performance among the three detection layers. The small-scale detection layer also performs better for small-sized wheat spikes than for large ones.

Particularly, the large-scale detection layer ignores 30% of small-sized wheat spikes with an average size of 3434 pixels and misclassifies background areas as wheat spikes. It has difficulty distinguishing spike/background areas in Fig. 6 where 67 wheat spikes are labeled.

### The performance of the improved network

In summary, the large-scale detection layer performs the worst, while the medium-scale and small-scale detection layers are relatively better. Therefore, the network structure is streamlined by removing the large-scale detection layer, enhancing feature fusion in the medium-scale detection layer, and adding a micro-scale detection layer to enhance the network's performance in detecting small-sized wheat spikes.

We compare the standard YOLOv5 and the improved network on the GWHD dataset (Fig. 7). The improved network increases AP by 0.5% compared to standard YOLOv5 and achieves the best AP of 93.5% on the $600 \times 600$ resolution image. The largest improvement is achieved in the $150 \times 150$ resolution image training, with an AP improvement of 7.4%. The wheat spike detection network can be improved based on the proposed interpretive analysis. Furthermore, although the proposed method results in a slight decrease in FPS to 130 and a slight increase in the number of network layers to 236, the network parameters are reduced from 7 to 6 M, and AP is improved in each input size (Table 5).

## Discussion

Scale issue has always been an important research problem in wheat spike detection, similar to other object detection tasks [28, 29]. The scale issue in the wheat spike detection network exists in terms of the multi-scale

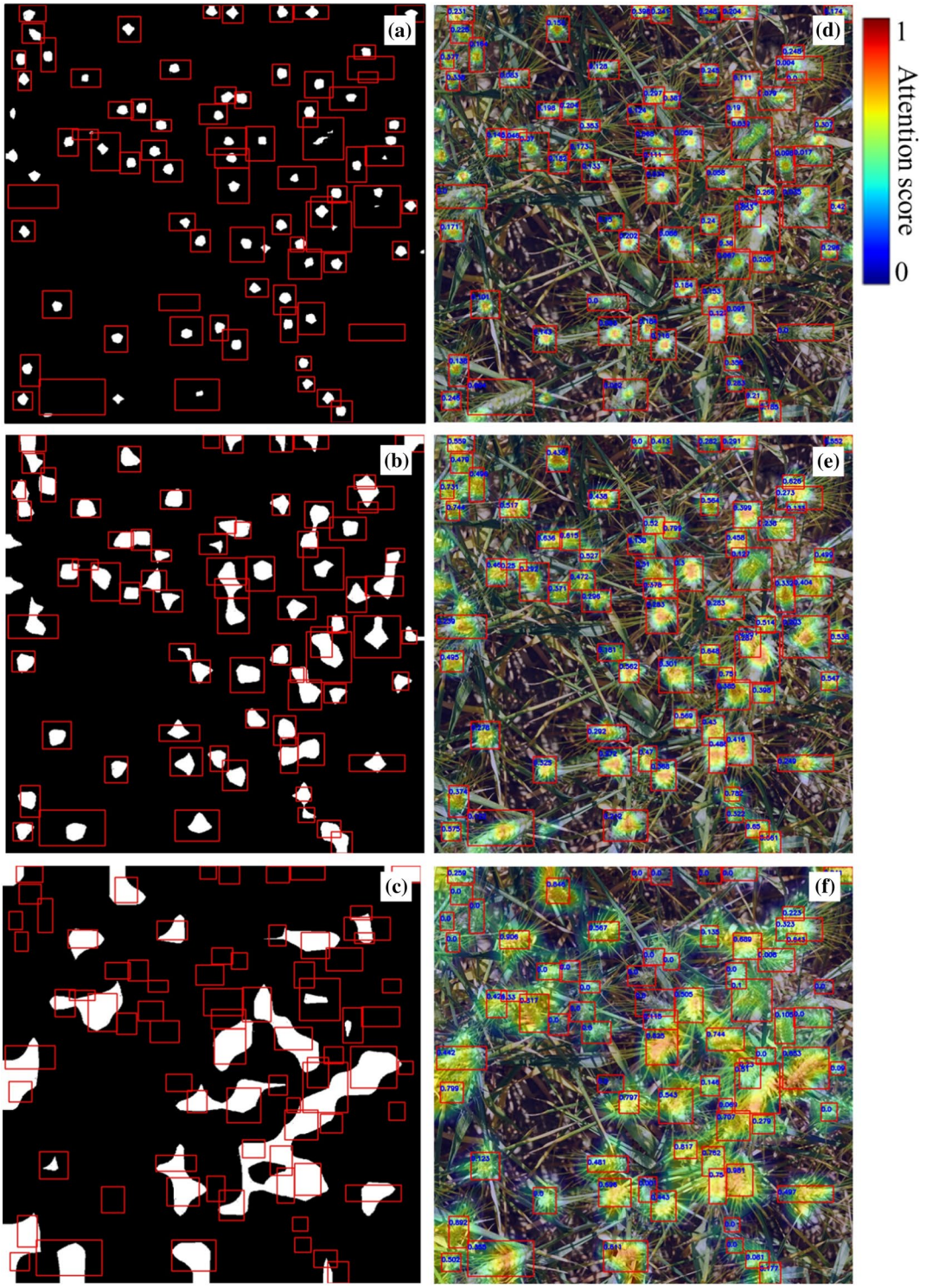


**Fig. 4** Attention areas and attention scores of multi-scale detection layers (red rectangles are the prior labeled wheat spike bounding boxes). **a** Attention areas of the small-scale detection layer (in white). **b** Attention areas of the medium-scale detection layer (in white). **c** Attention areas of the large-scale detection layer (in white). **d** Attention score values and heatmaps of the small-scale detection layer. **e** Attention score values and heatmaps of the medium-scale detection layer. **f** Attention score values and heatmaps of the large-scale detection layer

**Table 3** Attention score statistics of multi-scale detection layers

| Detection layer | Attention score | Proportion of spikes (%) | Average spike size (Pixels) | Total attention score range | Mean attention score |
|---|---|---|---|---|---|
| Small-scale detection layer | 0.0–0.1 | 64.4 | 8682 | 0–0.607 | 0.083 |
| | 0.1–0.2 | 27.5 | 3438 | | |
| | 0.2–0.3 | 6.3 | 2137 | | |
| | 0.3–0.4 | 1.4 | 1581 | | |
| | 0.4–0.5 | 0.3 | 1188 | | |
| | 0.5–0.6 | 0.07 | 961 | | |
| | 0.6–0.7 | 0.03 | 733 | | |
| | 0.7–0.8 | – | – | | |
| | 0.8–0.9 | – | – | | |
| | 0.9–1.0 | – | – | | |
| Medium-scale detection layer | 0.0–0.1 | 13.0 | 12030 | 0–0.984 | 0.25 |
| | 0.1–0.2 | 26.4 | 8949 | | |
| | 0.2–0.3 | 27.9 | 5619 | | |
| | 0.3–0.4 | 17.8 | 4165 | | |
| | 0.4–0.5 | 9.0 | 3360 | | |
| | 0.5–0.6 | 3.9 | 2758 | | |
| | 0.6–0.7 | 1.4 | 2347 | | |
| | 0.7–0.8 | 0.5 | 2003 | | |
| | 0.8–0.9 | 0.08 | 1742 | | |
| | 0.9–1.0 | 0.02 | 1527 | | |
| Large-scale detection layer | 0.0–0.1 | 38.8 | 3891 | 0–1 | 0.309 |
| | 0.1–0.2 | 6.3 | 10298 | | |
| | 0.2–0.3 | 7.3 | 11365 | | |
| | 0.3–0.4 | 8.7 | 10581 | | |
| | 0.4–0.5 | 9.0 | 8946 | | |
| | 0.5–0.6 | 9.2 | 7962 | | |
| | 0.6–0.7 | 8.2 | 6861 | | |
| | 0.7–0.8 | 6.6 | 6092 | | |
| | 0.8–0.9 | 4.0 | 5297 | | |
| | 0.9–1.0 | 1.9 | 4346 | | |

input images and the multi-scale of network layers. They cause an impact on the construction of the wheat spike detection network, including the model efficiency and performance. Therefore, it is necessary to carry out interpretive analysis and scale optimization of the network. Due to the limitations of different image acquisition platforms, there is an established problem that objects vary in size due to their physical morphology in covering datasets. The size of wheat spikes in images varies significantly in the study of wheat spike detection. Some researchers directly start from the size characteristics of objects in datasets and determine the optimal pixel size by up-sampling the small objects in original images. It effectively improves detection accuracy, but too many upsampling operations increase the processing time and lead to more false detections [30]. Based on the above, relevant studies deliberately select labeled datasets with significant differences in object scales [31] and sufficient data amount [32, 33] in the preliminary dataset preparation stage for detection. The proposed Feature Pyramid Network (FPN) solves the multi-scale object detection problem at the network structure level. The problem is successfully settled by building an FPN structure for multi-scale detection [34].

Adjusting network structure will affect object detection accuracy for a deep learning network [35]. Based on subjective experience, researchers have enhanced the detection network's performance by adding a micro-scale detection layer [36], adjusting feature enhancement modules [37–43], and rotating original horizontal detection boxes [44–46]. However, the studies mentioned above focused merely on the direct application of prior knowledge and thus lacked significant support from interpretive works [47].

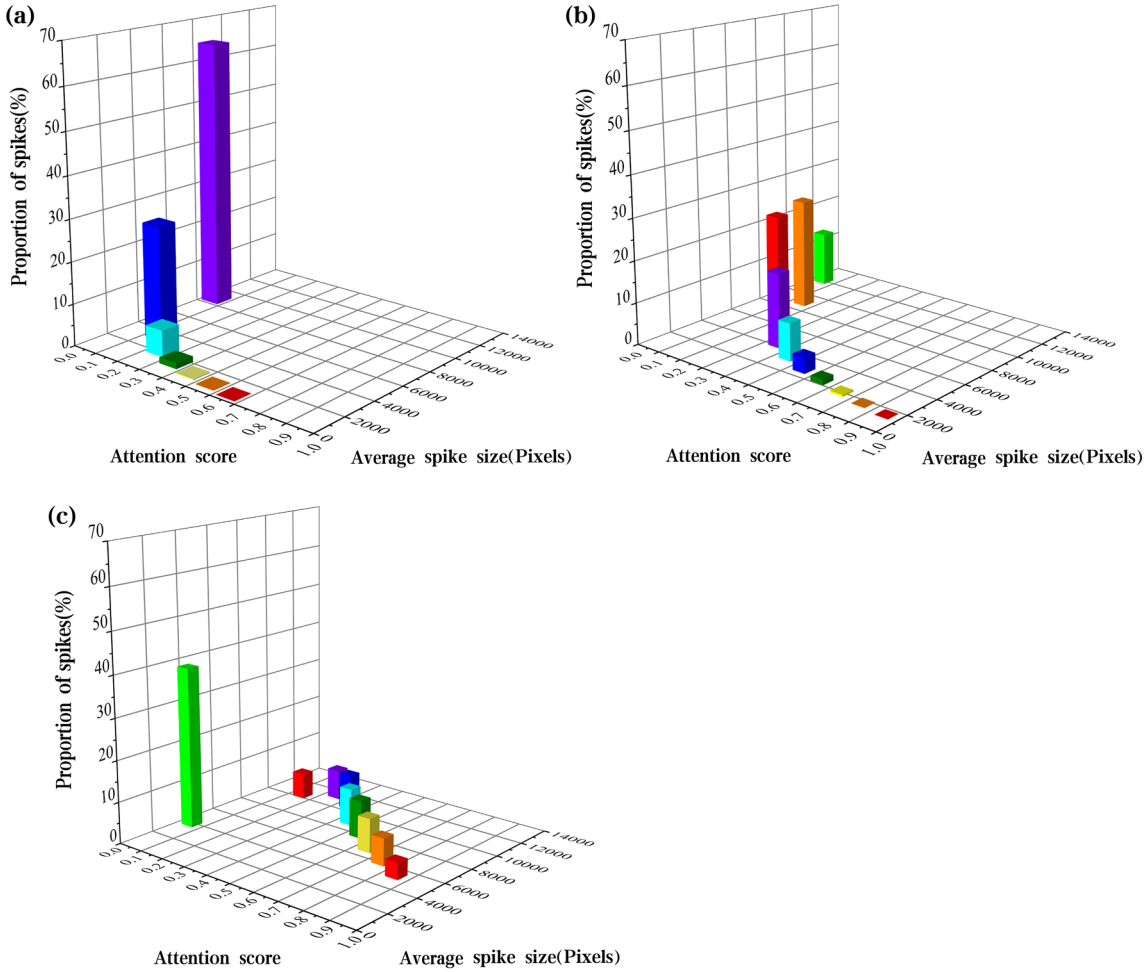


**Fig. 5** Spike distribution in different attention score ranges of three-scale detection layers. **a** Small-scale detection layer. **b** Medium-scale detection layer. **c** Large-scale detection layer

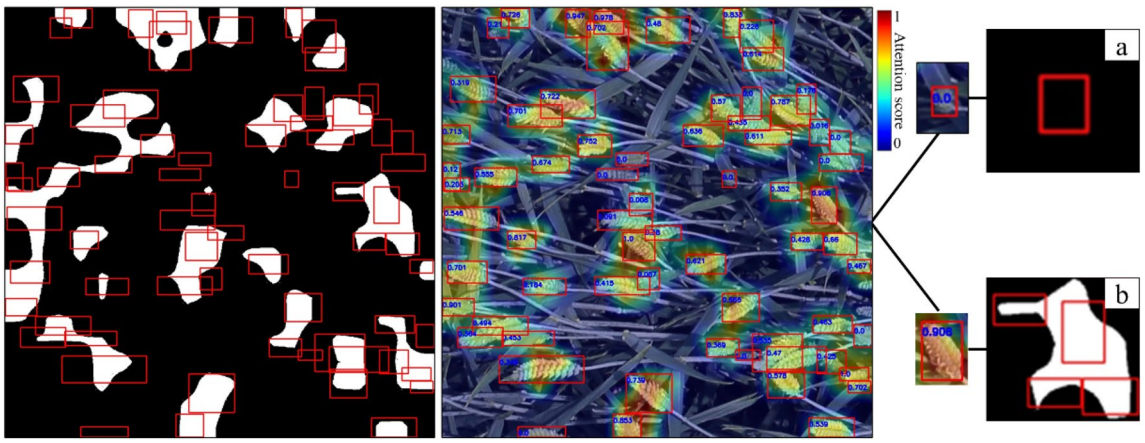


**Fig. 6** Two typical problems in the large-scale detection layer. **a** Absence of network attention for a small-sized wheat spike in this detection layer. **b** The attention area of this detection layer significantly confuses the wheat spike with the non-spike background region. Regions of white/black pixels are network attention areas and backgrounds separately.

**Table 4** Spikes with an attention score of 0 in three detection layers

| Detection layer | Mean spike size (Pixels) | Proportion of spikes (%) |
|---|---|---|
| Small-scale | 13040 | 17.5 |
| Medium-scale | 4660 | 4.0 |
| Large-scale | 3434 | 30.0 |

Most interpretive research provides qualitative explanations by outputting saliency maps of a network to provide a sound scientific basis for network refinement [48–50]. With saliency maps, researchers can visualize the location and size of network attention areas [51]. However, quantitative metrics are lacking in these studies for further network performance evaluation. In the proposed research, attention areas extracted from different scale detection layers show significant scale effects. They can accurately reflect semantic and location information of wheat spikes in each detection layer (Fig. 8). With the Grad-CAM algorithm, we successfully quantitatively describe the scale effects and provides a scientific basis for the scale optimization of the network.

It is visually evident that the attention area of small-scale and medium-scale detection layers accurately reflects wheat spikes' morphology and spatial location. Small-scale and medium-scale detection layers successfully focus on 82.5% and 96% of wheat spike objects in all 44538 wheat spikes. Attention areas of the large-scale detection layer are far beyond areas where wheat spikes locate. The large-scale detection layer focuses on merely 30% of wheat spikes, and there is confusion between wheat spike areas and background areas (Fig. 8). This confusion means that the wrong attention is paid to non-spike areas. It may be related to the receptive field of the neural network. The size in pixels of feature maps output

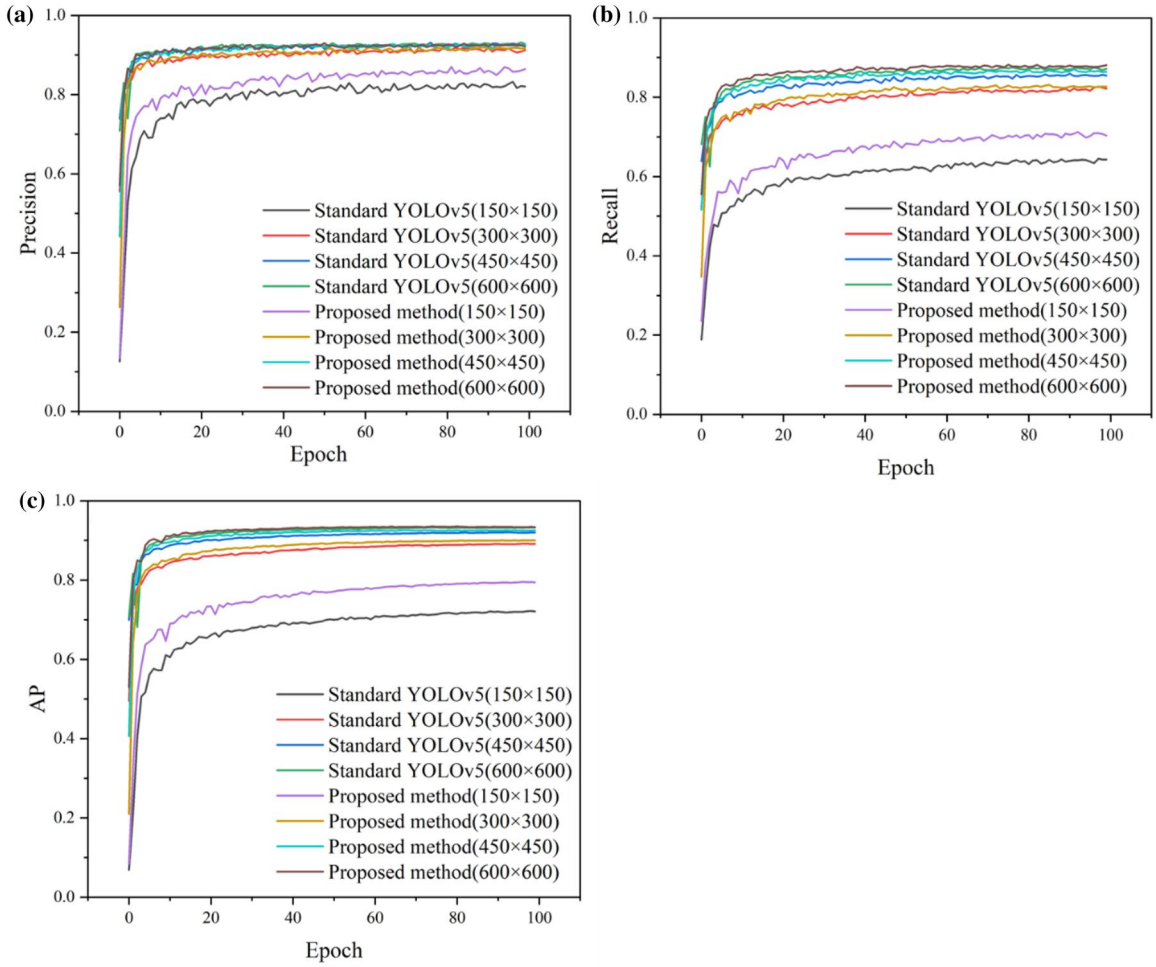


**Fig. 7** *Precision*, *Recall*, and *AP* curves of the wheat spike detection for the improved method and standard YOLOv5. **a** The precision curves. **b** The recall curves. **c** The *AP* curves

**Table 5** The performance of the improved network vs. standard YOLOv5

| Method | Input size | FPS | Layers | Parameters | AP (%) |
|---|---|---|---|---|---|
| Standard YOLOv5 | 150×150 | 166 | 213 | 7 M | 72.2 |
| | 300×300 | 166 | 213 | 7 M | 89.2 |
| | 450×450 | 166 | 213 | 7 M | 92.0 |
| | 600×600 | 166 | 213 | 7 M | 93.0 |
| Improved network | 150×150 | 130 | 236 | 6 M | 79.6 |
| | 300×300 | 130 | 236 | 6 M | 90.0 |
| | 450×450 | 130 | 236 | 6 M | 92.5 |
| | 600×600 | 130 | 236 | 6 M | 93.5 |

**Table 6** The performance of the standard YOLOv5 vs. YOLOv5[a]

| Method | Input size | FPS | Layers | Parameters | AP (%) |
|---|---|---|---|---|---|
| YOLOv5 | 150×150 | 166 | 213 | 7 M | 72.2 |
| | 300×300 | 166 | 213 | 7 M | 89.2 |
| | 450×450 | 166 | 213 | 7 M | 92.0 |
| | 600×600 | 166 | 213 | 7 M | 93.0 |
| YOLOv5[a] | 150×150 | 188 | 190 | 5 M | 74.9 |
| | 300×300 | 188 | 190 | 5 M | 89.4 |
| | 450×450 | 188 | 190 | 5 M | 92.5 |
| | 600×600 | 188 | 190 | 5 M | 93.3 |

[a] YOLOv5[a] represents the standard YOLOv5 without a large-scale detection layer

by each detection layer decreases exponentially with a factor of 4 from the small-scale to the large-scale detection layer. Meanwhile, the corresponding size in pixels of receptive fields increases exponentially with a factor of 4 [52, 53]. It is consistent with the situation presented in the graph (Fig. 8).

Moreover, according to interpretive analysis, the large-scale detection layer performs poorly in detecting wheat spikes in the GWHD dataset. The new network without a large-scale detection layer achieves overall improvements in all result metrics in multi-resolution jobs (Table 6). This finding is consistent with other research results achieving better results in higher resolution training [54].

This study aims to explore the combination of wheat spike features and interpretability methods to construct a wheat spike detection network. This is a general improvement method that can be applied to various single-stage object detection models, including YOLOv5, YOLOv6 and the latest YOLOv7. Existing object detection models are evolving towards large-scale and universal models with massive parameters, making training difficult and leading to high computational costs [55, 56]. This paper integrates interpretable methods to construct and optimize a wheat spike detection model for complex scenes without too many parameters, providing theoretical foundations for model development (Table 5).

The study has only carried out interpretive research on three-scale detection layers and conducted scale refinement for these layers. In future work, it will be meaningful to introduce attention-based interpretive work on the network's backbone module to explore its improvement path. We also plan to further explain how the convolutional layers and kernels in the neural network affect the accuracy of wheat spike detection. Meanwhile, more diverse wheat spike datasets are needed to validate our method to ensure a convincing and objective research finding.

## Conclusion

The study proposes a scale refinement method for the detection layers of the wheat spike detection network based on the deep learning interpretive method Grad-CAM. A more streamlined wheat spike detection network is successfully constructed and performs well on the GWHD dataset with better detection accuracy and lower model complexity. Compared to previous work, our study has two novel aspects. First, the proposed method integrates features with prior knowledge without directly referencing and superimposing novel technologies in

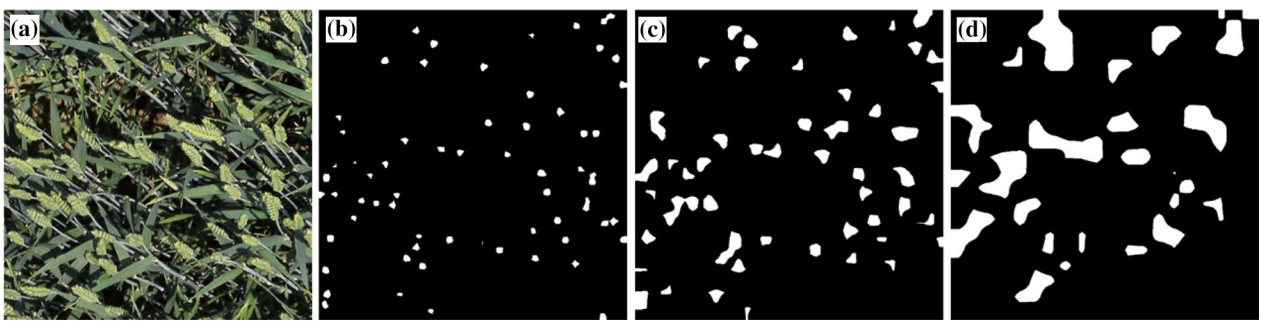


**Fig. 8** Attention area of different scale detection layers (in white): **a** Original wheat spike image. **b** Attention area of the small-scale detection layer. **c** Attention area of the medium-scale detection layer. **d** Attention area of the large-scale detection layer

object detection. By analyzing the size features of wheat spikes, we design a superior wheat spike detection network. Second, we demonstrate the effectiveness of the improved modules from both theoretical and experimental perspectives. The size characteristics of wheat spikes in the dataset are quantitatively analyzed and the results are used to optimize the wheat spike detection network. The study provides a new theoretical basis for research on wheat spike detection based on deep learning. It offers a technical reference for constructing and developing wheat spike detection networks with better robustness, generality, and applicability.

**Acknowledgements**
This research was supported by the National Natural Science Foundation of China (Grant No. 32171892)

**Author contributions**
JY performed experiments, analyzed the data, and wrote the manuscript. JZ performed experiments, analyzed the data, and wrote the manuscript. YC performed experiments and prepared data visualization. SW performed experiments. XQ developed software used in this work. XY made provisions for study materials. YT supervised the research activity planning and execution. YZ supervised the research activity planning and execution. WC managed and coordinated the research activity planning and execution. XZ conceived the research, guided the entire study, revised the manuscript, and provided valuable comments and suggestions. All the authors approved the manuscript and have made all required statements and declarations. All authors read and approved the final manuscript.

**Availability of data and materials**
The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 16 January 2023 Accepted: 29 April 2023
Published online: 13 May 2023

**References**

1. FAOSTAT. http://www.fao.org/faostat/en/. Accessed 22 Dec 2022.
2. Hasan MM, Chopin JP, Laga H, Miklavcic SJ. Detection and analysis of wheat spikes using convolutional neural networks. Plant Methods. 2018;14(1):1–13. https://doi.org/10.1186/s13007-018-0366-8.
3. Thakur AK, Singh S, Goyal N, Gupta K. A comparative analysis on the existing techniques of wheat spike detection. In: 2021 2nd International Conference for Emerging Technology (INCET). IEEE. 2021. pp. 1–6. https://doi.org/10.1109/INCET51464.2021.9456284
4. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. pp. 580–7. https://doi.org/10.1109/CVPR.2014.81.
5. Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). 2015. pp. 1440–8. https://doi.org/10.1109/ICCV.2015.169.
6. 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(06):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
7. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2016. pp. 779–88. https://doi.org/10.1109/CVPR.2016.91.
8. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society. 2017. pp. 6517–25. https://doi.org/10.1109/CVPR.2017.690.
9. Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv. 2018. https://doi.org/10.48550/arXiv.1804.02767.
10. Bochkovskiy A, Wang CY, Liao HYM. Yolov4: optimal speed and accuracy of object detection. arXiv. 2020. https://doi.org/10.48550/arXiv.2004.10934.
11. Ultralytics. YOLOv5. https://github.com/ultralytics/yolov5. Accessed 1 Mar 2022.
12. Yang B, Gao Z, Gao Y, Zhu Y. Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module. Agronomy. 2021;11(6):1202. https://doi.org/10.3390/agronomy11061202.
13. Bhagat S, Kokare M, Haswani V, Hambarde P, Kamble R. WheatNet-Lite: a novel light weight network for wheat head detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE. 2021. pp. 1332–41. https://doi.org/10.1109/ICCVW54120.2021.00154.
14. Wang Y, Qin Y, Cui J. Occlusion robust wheat ear counting algorithm based on deep learning. Front Plant Sci. 2021;12:645899. https://doi.org/10.3389/fpls.2021.645899.
15. Gong B, Ergu D, Cai Y, Ma B. Real-time detection for wheat head applying deep neural network. Sensors. 2020;21(1):191. https://doi.org/10.3390/s21010191.
16. Fernandez-Gallego JA, Kefauver SC, Gutiérrez NA, Nieto-Taladriz MT, Araus JL. Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. Plant Methods. 2018;14:1–12. https://doi.org/10.1186/s13007-018-0289-4.
17. Zhu Y, Cao Z, Lu H, Li Y, Xiao Y. In-field automatic observation of wheat heading stage using computer vision. Biosys Eng. 2016;143:28–41. https://doi.org/10.1016/j.biosystemseng.2015.12.015.
18. Xiang Yu, Choi W, Lin Y, Savarese S. Subcategory-aware convolutional neural networks for object proposals and detection. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE. 2017. pp. 924–33. https://doi.org/10.1109/WACV.2017.108.
19. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE. 2012. pp. 3354–61. https://doi.org/10.1109/CVPR.2012.6248074
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vision. 2020;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7.
21. David E, Madec S, Sadeghi-Tehran P, Aasen H, Zheng B, Liu S, Kirchgessner N, Ishikawa G, Nagasawa K, Badhon MA, Pozniak C, Solan B, Hund A, Chapman SC, Baret F, Stavness I, Guo W. Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. Plant Phenom. 2020. https://doi.org/10.34133/2020/3521852.
22. David E, Serouart M, Smith D, Madec S, Velumani K, Liu S, Wang X, Pinto F, Shafiee S, Tahir ISA, Tsujimoto H, Nasuda S, Zheng B, Kirchgessner N, Aasen H, Hund A, Sadhegi-Tehran P, Nagasawa K, Ishikawa G, Dandrifosse S, Carlier A, Dumont B, Mercatoris B, Evers B, Kuroki K, Wang H, Ishii M, Badhon MA, Pozniak C, LeBauer DS, Lillemo M, Poland J, Chapman S, Solan B, Baret F, Stavness I, Guo W. Global wheat head detection 2021: an improved dataset for benchmarking wheat head detection methods. Plant Phenom. 2021. https://doi.org/10.34133/2021/9846158.
23. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2018. pp. 8759–68. https://doi.org/10.1109/CVPR.2018.00913.

24. Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. pp. 10781–90. https://doi.org/10.1109/CVPR.2018.00913.
25. Fan FL, Xiong J, Li M, Wang G. On interpretability of artificial neural networks: a survey. IEEE Trans Radiat Plasma Med Sci. 2021;5(6):741–60. https://doi.org/10.1109/TRPMS.2021.3066428.
26. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2016. pp. 2921–9. https://doi.org/10.1109/CVPR.2016.319.
27. Zhao J, Zhang X, Yan J, Qiu X, Yao X, Tian Y, Zhu Y, Cao W. A wheat spike detection method in UAV images based on improved YOLOv5. Remote Sens. 2021;13(16):3095. https://doi.org/10.3390/rs13163095.
28. Xiao Y, Tian Z, Yu J, Zhang Y, Liu S, Du S, Lan X. A review of object detection based on deep learning. Multimed Tools Appl. 2020;79(33):23729–91. https://doi.org/10.1007/s11042-020-08976-6.
29. Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J. Detnet: design backbone for object detection. In: European Conference on Computer Vision. Cham: Springer. 2018. pp. 339–54. https://doi.org/10.1007/978-3-030-01240-3_21.
30. Mansour A, Hussein W M, Said E. Small objects detection in satellite images using deep learning. In: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE. 2019. pp. 86–91. https://doi.org/10.1109/ICICIS46948.2019.9014842.
31. Pang Y, Cao J, Li Y, Xie J, Sun H, Gong J. TJU-DHD: a diverse high-resolution dataset for object detection. IEEE Trans Image Process. 2020;30:207–19. https://doi.org/10.1109/TIP.2020.3034487.
32. Duan R, Deng H, Tian M, Deng Y, Lin J. SODA: a large-scale open site object detection dataset for deep learning in construction. Autom Constr. 2022;142:104499. https://doi.org/10.1016/j.autcon.2022.104499.
33. Pathak AR, Pandey M, Rautaray S. Application of deep learning for object detection. Procedia Comput Sci. 2018;132:1706–17. https://doi.org/10.1016/j.procs.2018.05.144.
34. Lin T Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2017. pp. 936–44. https://doi.org/10.1109/CVPR.2017.106.
35. Woo S, Park J, Lee J Y, Kweon I S. CBAM: convolutional block attention module. In: European Conference on Computer Vision. Cham: Springer. 2018. pp. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
36. Zhang R, Wen C. SOD-YOLO: a small target defect detection algorithm for wind turbine blades based on improved YOLOv5. Adv Theory Simul. 2022. https://doi.org/10.1002/adts.202100631.
37. Qi G, Zhang Y, Wang K, Mazur N, Liu Y, Malaviya D. Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion. Remote Sens. 2022;14(2):420. https://doi.org/10.3390/rs14020420.
38. Gong Y, Yu X, Ding Y, Peng X, Zhao J, Han Z. Effective fusion factor in FPN for tiny object detection. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE. 2021. pp. 1159–67. https://doi.org/10.1109/WACV48630.2021.00120.
39. Jing Y, Ren Y, Liu Y, Wang D, Yu L. Automatic extraction of damaged houses by earthquake based on improved YOLOv5: a case study in Yangbi. Remote Sens. 2022;14(2):382. https://doi.org/10.3390/rs14020382.
40. Sun Z, Yang H, Zhang Z, Liu J, Zhang X. An improved YOLOv5-based tapping trajectory detection method for natural rubber trees. Agriculture. 2022;12(9):1309. https://doi.org/10.3390/agriculture12091309.
41. Liao X, Lv S, Li D, Luo Y, Zhu Z, Jiang C. YOLOv4-MN3 for PCB surface defect detection. Appl Sci. 2021;11(24):11701. https://doi.org/10.3390/app112411701.
42. Deng Z, Sun H, Zhou S, Zhao J, Lei L, Zou H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J Photogr Remote Sens. 2018;145:3–22. https://doi.org/10.1016/j.isprsjprs.2018.04.003.
43. Liu B, Luo H. An improved Yolov5 for multi-rotor UAV detection. Electronics. 2022;11(15):2330. https://doi.org/10.3390/electronics11152330.
44. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K. SCRDet: towards more robust detection for small, cluttered and rotated objects. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE. 2019. pp 8231–40. https://doi.org/10.1109/ICCV.2019.00832.
45. Chen C, Zhong J, Tan Y. Multiple-oriented and small object detection with convolutional neural networks for aerial image. Remote Sens. 2019;11(18):2176. https://doi.org/10.3390/rs11182176.
46. Zhao J, Yan J, Xue T, Wang S, Qiu X, Yao X, Tian Y, Zhu Y, Cao W, Zhang X. A deep learning method for oriented and small wheat spike detection (OSWSDet) in UAV images. Comput Electron Agric. 2022;198:107087. https://doi.org/10.1016/j.compag.2022.107087.
47. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):1–42. https://doi.org/10.1145/3236009.
48. Ghose D, Desai SM, Bhattacharya S, Chakraborty D, Fiterau M, Rahman T. Pedestrian detection in thermal images using saliency maps. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2019. pp. 988–97. https://doi.org/10.1109/CVPRW.2019.00130.
49. Brahimi M, Arsenovic M, Laraba S, Sladojevic S, Boukhalfa K, Moussaoui A. Deep learning for plant diseases: detection and saliency map visualization. Human and machine learning. Cham: Springer; 2018. https://doi.org/10.1007/978-3-319-90403-0_6.
50. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Li MD, Kalpathy-Cramer J. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell. 2021. https://doi.org/10.1148/ryai.2021200267.
51. Nagasubramanian K, Singh AK, Singh A, Sarkar S, Ganapathysubramanian B. Usefulness of interpretability methods to explain deep learning based plant stress phenotyping. arXiv. 2020. https://doi.org/10.48550/arXiv.2007.05729.
52. Zhang Y, Shen T. Small object detection with multiple receptive fields. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing. 2020; 440(3): 032093. https://doi.org/10.1088/1755-1315/440/3/032093.
53. Cao J, Chen Q, Guo J, Shi R. Attention-guided context feature pyramid network for object detection. arXiv. 2020. https://doi.org/10.48550/arXiv.2005.11475.
54. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiol Artif Intell. 2020. https://doi.org/10.1148/ryai.2019190015.
55. Carion N, Massa F, Synnaeve G. End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020. pp. 213–29. https://doi.org/10.1007/978-3-030-58452-8_13
56. Wang D, Zhang J, Du B, et al. An empirical study of remote sensing pretraining. IEEE Trans Geosci Remote Sens. 2022. https://doi.org/10.1109/LGRS.2022.3143368.


## Publisher's Note