SEMI-SUPERVISED WHEAT EAR DETECTION ALGORITHM BASED ON THE MODIFIED YOLOv8
/
基于改进 YOLOv8 的半监督麦穗识别算法

Yu ZHANG [1)], Zhihui XU [1)], Fuzhong LI [1)], Xiaoying ZHANG *[1)], Xiao CUI *[1)]
[1)] School of Software, Shanxi Agricultural University, Shanxi/China;
Tel: +86-15803449361; E-mail: xiaoyingzhang@sxau.edu.cn; cuixiaocuilu@163.com
DOI: https://doi.org/10.35633/inmateh-75-54

Keywords: *wheat spike, object detection, semi-supervised learning, YOLOv8*

ABSTRACT

In contemporary agricultural practices, the use of image and video acquisition technologies, such as drones and cameras, has become increasingly common for capturing and monitoring crop growth in agricultural fields. The reliance on visual data for analyzing farm management conditions and facilitating decision-making processes is gaining significant traction. However, in practical applications, image acquisition tools often face challenges in maintaining optimal distance and angle during data capture, which can negatively impact the detection accuracy of existing object detection methods. Semi-supervised learning plays a crucial role in improving object detection. In this study, a semi-supervised algorithm for wheat spike recognition was developed based on an optimized YOLOv8n model. The model incorporates SPDConv and PSA attention modules after the SPPF layer, effectively reducing computational and memory demands while enhancing model performance. The proposed model achieved an accuracy of 94.2%, outperforming YOLOv5s, Efficient Teacher, and the baseline YOLOv8n by 10.9%, 4.5%, and 6.1%, respectively—demonstrating its strong potential for practical agricultural applications.

摘要

当今农业，使用无人机或摄像头这样的图像视频采集工具拍摄或监控农田作物生长情况变得常见，依靠这些图像分析农田管理情况和辅助决策也会变得越来越有市场。但在实际生产中，图像采集工具在拍摄时很难保持合理的距离和角度，这使得现有的目标检测方法检测精度低。而半监督学习对提高物体检测至关重要。我们开发了一种用于麦穗识别的半监督学习算法，使用了改进的 YOLOv8n，并在 SPPF 之后集成了 PSA 注意力机制和 SPDConv。这些改进减少了计算和内存需求，提升了模型性能。我们的模型达到了 94.2%的准确率，分别比 YOLOv5s、Efficient Teacher 和 YOLOv8n 高出 10.9%、4.5%和 6.1%，展示了其实际潜力。

INTRODUCTION

Wheat, recognized as a globally cultivated food crop, has maintained its status as the third most extensively sown crop in terms of both area and total production since the 1940s. It serves as a staple food for approximately one-third of the global population and is cultivated across an estimated 224 million hectares, thereby supporting around 30 percent of the world's inhabitants *(Eversole et al., 2014; Li et al., 2018a)*. Enhancing wheat yield per unit area remains a primary objective of contemporary breeding efforts. The implementation of automated detection and counting of wheat ears facilitates the rapid and precise assessment of ear numbers, which subsequently allows for the estimation of yield per plant, yield per unit area, and overall production. This advancement significantly aids researchers in optimizing breeding efficiency. With continuous advancements in artificial intelligence technologies, various approaches, including image processing, machine learning, and deep learning, are increasingly being applied to wheat yield prediction and phenotypic identification.

The methodologies employed for counting wheat ears through traditional image processing techniques can be broadly classified into three distinct categories *(Fu Jingbo, 2021; Liu et al., 2019)*: curve fitting segmentation, image-based fractal segmentation, and pixel-based area estimation *(Liu Tao, 2014)*. These approaches typically utilize color features or grain characteristics to extract images of wheat ears, followed by morphological operations such as corrosion expansion, cavity filling, and refinement processing. This sequence of operations facilitates the enumeration of wheat spike skeleton images through angular point detection methods. Moreover, *Li et al., (2018b),* implemented a technique that involved color space

transformation, utilizing a custom device to capture RGB images of wheat in natural field conditions, which were subsequently converted to HSV space.

The saturation component (S) was subjected to binary conversion, resulting in a binary image. An adhesion object segmentation algorithm was then applied to the wheat ear images, followed by the implementation of a concave point detection method for segmentation and counting, ultimately enabling the calculation of grain numbers and yield prediction. Furthermore, *Fernandez-Gallego et al., (2018),* introduced a wheat ear counting method based on an enhanced K-means clustering algorithm, which relies on color feature clustering. In this approach, each sub-region identified through clustering is interpreted as representing a wheat ear, with the total number of sub-regions serving as an estimate of the wheat ear count. While traditional machine learning techniques for wheat ear counting significantly reduce the labor required compared to manual counting and enhance efficiency and accuracy to some extent, the complexity of field environments presents challenges in maintaining consistent image quality and limits the versatility of these methods.

The continuous advancements in deep learning within the domain of agricultural research have prompted certain scholars to integrate YOLOv5 with semi-supervised learning methodologies, culminating in the creation of a semi-supervised target detection algorithm that is applicable across various contexts. In their study, *Zhou et al., (2023),* presented an innovative semi-supervised adaptive algorithm, named SSDA-YOLO, which combines a YOLOv5-based semi-supervised framework with a knowledge distillation approach. This algorithm enhances the quality of pseudo-labels through improved view refinement and global view filtering. Furthermore, *Lyu et al., (2022),* reported a notable enhancement in the recognition accuracy of citrus bagging by incorporating a strip attention module into the YOLOv5 backbone and employing additional semi-supervised learning techniques.

Despite the advancements made in the aforementioned studies, which have enhanced the algorithm from various perspectives and yielded certain outcomes, significant challenges persist when applying these methods to single-stage target detection models, such as YOLO. In particular, the detection of wheat ears presents several critical issues. Firstly, there is an inconsistency in the scale of wheat spike targets. As wheat ears progress from the heading stage to maturity throughout their growth cycle, the size of individual wheat plants varies temporally. Additionally, certain weeds exhibit morphological similarities to wheat ears, thereby complicating the accurate differentiation and identification of these targets based solely on visual characteristics. Secondly, the samples of wheat spikes are frequently densely clustered. In areas where wheat plants are closely spaced, occlusion of wheat ears is a prevalent issue. The dataset contains a substantial number of samples, which not only affects the accuracy of target identification but also complicates the labeling process. Finally, the extensive volume of data within the wheat spike dataset results in high labor costs associated with large-scale annotation.

In response to the challenges previously outlined, this study proposes a semi-supervised algorithm for the detection of wheat spikes, which is based on an enhanced version of YOLOv8. The algorithm integrates Spatial Depth Conversion Convolution (SPDConv) to improve the detection of small objects in low-resolution images, thereby augmenting the feature extraction capabilities for such targets *(Sunkara and Luo, 2023)*. Additionally, the PSA attention mechanism is incorporated to reduce the computational and memory demands associated with self-attention in visual tasks, ultimately resulting in enhanced detection accuracy. Furthermore, this paper presents improvements derived from the modified Efficient Teacher framework *(Xu et al., 2023)*, which effectively increases the detection accuracy of the algorithm by utilizing valuable information from unlabeled images, all while preserving the model's original size.

MATERIALS AND METHODS

Data collection and data processing

To enhance the generalization performance of the model, this study employed two distinct datasets for training. One of these datasets, known as the Wheat Detect (WD) dataset, consists of images of wheat captured at the Yang Jia Zhuang Village experimental base of Shanxi Agricultural University. The specific wheat variety used for this dataset is Agricultural University 212. The images were systematically collected from April to June 2024, with a frequency of every three days, to comprehensively cover the filling and maturation stages of the wheat.

The shooting process involved a variety of weather conditions, encompassing both sunny and overcast scenarios. The iPhone 15 Pro Max was employed as the primary photographic equipment, leading to the

acquisition of over 6,500 images of wheat canopies, of which 3,011 were selected for further analysis. All images were stored in JPG format, ensuring a consistent resolution of 3648×2736 pixels.

The LabelImg software was utilized for image annotation, with the sole category label assigned being "wheat." The resulting annotation data were subsequently saved as "txt" files in the YOLO format.

The second data set is the Global Wheat Head Detection (GWHD) from the global wheat public database *(David et al., 2021).* This data set pooled 6,422 RGB images with a resolution of 1,0,24×1,024 pixels, and 275,187 labeled ears of wheat. The images were collected from Europe, North America and Asia, covering a variety of varieties, planting conditions, climate types, as well as collection methods, thus ensuring the diversity of the data set in terms of genotype and environment. These characteristics of GWHD data sets help improving the accuracy and reliability of wheat ear detection and positioning. Fig.1 shows some examples of images generated from data obtained from the Yang Jia Zhuang Experimental Base.

Fig. 1 - Images of wheat at different growth stages

YOLOv8ps structure

As depicted in Fig.2, Efficient Teacher, which is constructed on YOLOv5, introduces several enhancements to the network. Efficient Teacher proposes three modules to implement a scalable and effective SSOD framework. The Dense Detector module improves the quality of pseudo labels with dense input and offers better inference efficiency; the Pseudo Label Assigner module categorizes pseudo labels into two types to mitigate the issue of pseudo labeling.

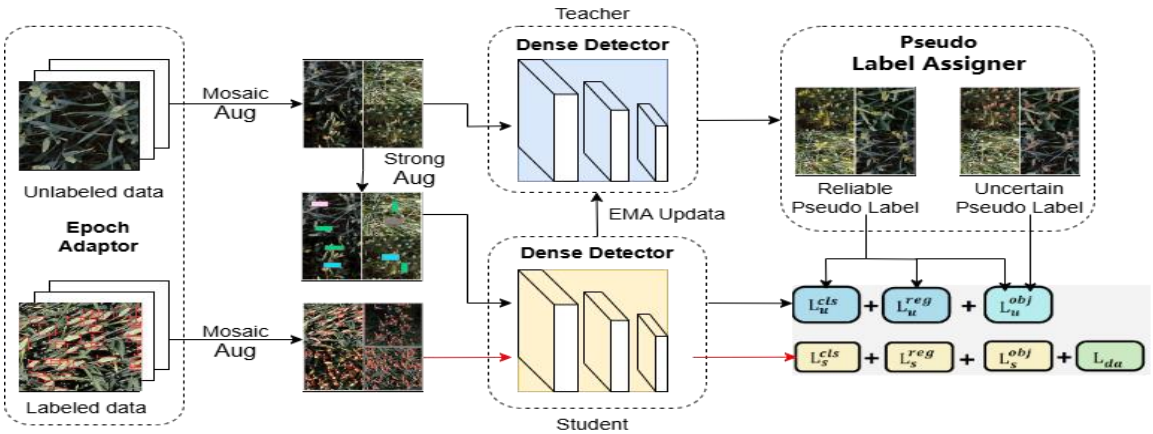

Fig. 2 - Semi-supervised training framework

Inconsistency problem, Epoch Adaptor reduces training time and the inconsistency of features. Although Efficient Teacher shows significant improvement in detection accuracy compared to SSOD and one-stage anchor-based detectors, it still faces significant challenges in the accuracy of small object detection. These challenges mainly stem from the fact that small objects occupy fewer pixels in images, making feature extraction difficult and thus affecting the accuracy of detection. To address these challenges, this study proposes a small object detection network architecture based on YOLOv8. The main innovations include: in the initial stage of the study, the YOLOv8ps model was proposed. By adopting this strategy, it was expected to capture the detailed features of small objects more accurately, thereby improving the overall detection performance. Subsequently, this study delved in to the efficiency of convolution (SPDConv) within CNN architectures and introduced four layers of SPDConv after the Conv layers in the YOLOv8 backbone network.

This method effectively maintains the detail information of the input image by transforming the spatial dimensions into depth dimensions, significantly enhancing the feature extraction performance for small objects.

Furthermore, to reduce the computational burden on the YOLOv8n network, Partial Self-Attention (PSA) was integrated, as proposed in YOLOv10. PSA addresses the high computational complexity and memory consumption associated with traditional self-attention mechanisms in visual tasks, thereby significantly enhancing the model's performance and capability.

As shown in Fig 3, the lightweight network architecture proposed in this study mainly consists of three core components: the backbone network, the neck structure, and the head module. The primary function of the backbone network is to extract feature information from the input image data and output feature maps at three different scales. During this feature extraction process, the study introduces the SPDConv module, which aims to take the feature maps produced by the previous convolution operation as input, process them through the SPD layer, and then perform further feature extraction through consecutive convolution layers. Additionally, the study integrates the PSA attention mechanism to address the inherent high computational complexity and significant memory consumption issues of traditional self-attention mechanisms in visual tasks.

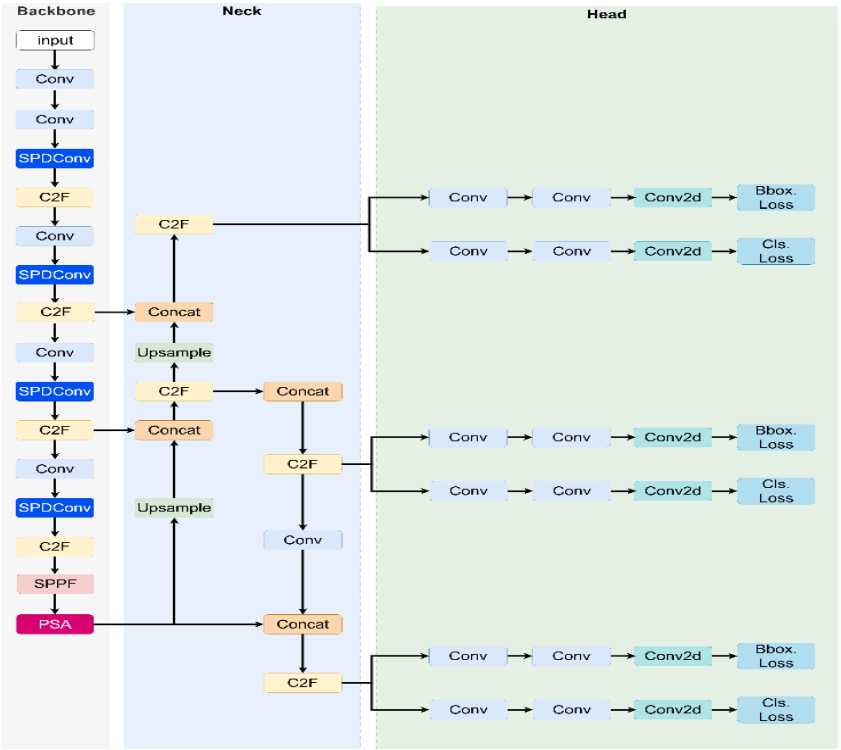

Fig. 3 - Network structure of improved YOLOv8 algorithm

The neck design of the network is used to enhance the ability to detect objects of different sizes by fusing feature maps from the backbone network at different scales, and then outputs three different scale feature maps. Similarly, both PSA and SPDConv are utilized for feature extraction and downsampling. Three independent and decoupled header structures are designed at the network's front, each responsible for predicting and analyzing feature maps of three distinct scales. This design allows each header structure to concentrate on processing feature information of a particular scale, thereby enhancing the network's capacity to capture and predict features across various scales. Consequently, this significantly improves the model's performance in visual tasks involving small objects. The darkened section in Fig. 3 represents the structural improvement module.

SPDConv The structure is composed of space to depth (SPD) layer and no convolutional step size (Conv) layer, replacing the pooling operation and convolution operation with step size in a similar way, and using scientific parameters to reduce the number of channels, which is suitable for a variety of convolutional neural network (CNN) architectures. In the traditional CNN architecture, the direct application of step convolution and pooling layer will lead to the gradual decrease of the spatial resolution of images with the deepening of the deepening of the network level, resulting in the loss of details of small objects, which makes the network encounter difficulties in accurately identifying these small objects. However, the combination of using SPD layers with step-free convolution layers enables the CNN to more effectively handle the challenges

posed by small objects and low-resolution images, thus improving the performance and robustness of the model in these complex scenarios.

For a specific feature plot X, the down sampling was performed by the scaling factor. Fig.4b shows the space-to-depth operation, where the spatial information is reorganized into the depth channel. When the scale factor is set to 2, the four feature subgraphs shown in Fig.4c can be obtained, and the shape of each subgraph is (S/2, S/2, C), which realizes the effect of subsampling the feature graph X by 2 times. Subsequently, the four feature sub-maps are successively connected along the channel dimension to form the feature map in Fig.4d (S/2, S/2,4C). After completing the feature transformation in the SPD layer, the dimensions change from Fig. 4d (D. 4C) to (S/2, S/2, D). The detailed procedure is shown in Figure Fig.4.

In this study, four SPDConv layers were introduced after the Conv convolution layer. This move aims to take the feature graph generated by the convolution operation of the previous layer as input, and to transform the spatial dimension of the input image into depth dimension, so as to improve the depth of the feature graph without losing information. Subsequently, the convolution processing through the continuous Conv layer realizes the feature extraction without reducing the feature graph size, which effectively maintains the detailed information of the image. Compared with the traditional convolution operation, this insertion method is more efficient in retaining the information in the channel, which significantly enhances the feature extraction performance for small targets (Fig 3).

Despite the self-attention mechanism, its computational amount and memory footprint are high. To solve this problem, YOLOv10 proposed an efficient local self-attention mechanism*(Wang et al., 2024)*, namely the PSA module. As shown in Fig.3, this module first divides the feature graph into two parts by channels after the convolution operation, one of which is sent to the N_PSA module composed of multi-head self-attention module (MHSA) and feedforward network (FFN). Subsequently, the two parts are reconnected and fused through a convolution operation. The PSA module is only placed after the lowest resolution stage, effectively avoiding the excessive consumption of resources caused by the squared computational complexity of the self-attention mechanism. Through this design, the PSA module can solve the high computational complexity and memory footprint problems faced by the self-attention mechanism in visual tasks, thus significantly improving the performance and capability of the model (Fig 5).

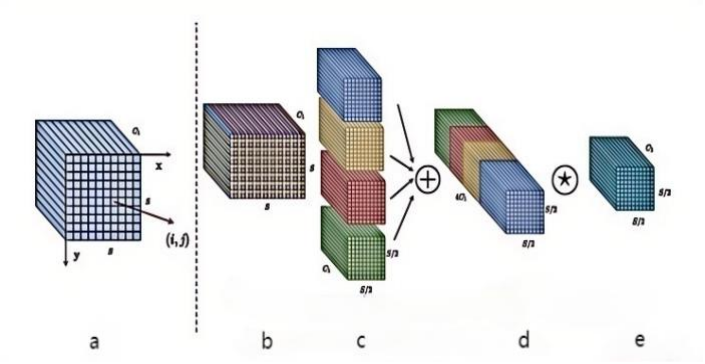

Fig. 4 - Illustration of SPDConv with a Scale Factor of 2

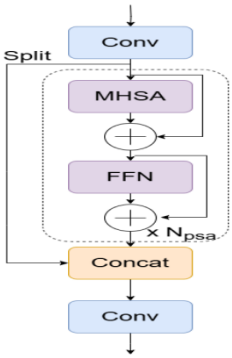

Fig. 5 - The partial self-Attention module (PSA)

Experimental Environment

The operating system used in this experiment is Windows 11, and the development environment of the model is Python 3.8, CUDA 11.1.0 and PyTorch 1.11.0. The training process was performed on the NVIDIA GeForce RTX 4060 graphics card. During the training process, YOLOv8n was used as the basic model, the size of the input image was set to 640×640 pixels, the batch size was set to 32, and the weight attenuation coefficient was set to 0.0005. In the fully supervised learning mode, the initial learning rate is set to 0.01, the minimum learning rate is set to 0.002, and the adjustment strategy of the learning rate adopts the cosine annealing method for a total of 40 training cycles (epochs). In the semi-supervised learning mode, the learning rate maintained a fixed value of 0.01, with a total of 300 training cycles.

Evaluation indexes of the model

A range of accepted evaluation measures in the field of target detection were used, including precision (Precision, P), recall (Recall, R), mean precision (mean Average Precision, mAP50) mAP50 at the IoU threshold of 0.5. Determine the integral of the precision-recall (P-R) curve.

In this study, only a single category of wheat ears was involved, with a sample size of n=1. Precision (P) is defined as the ratio of the number of correctly identified wheat ear samples to the total number of predicted

wheat ear samples, while Recall (R) is the ratio of the number of correctly identified wheat ear samples to the total number of actual wheat ear samples.

Model Detection Accuracy Analysis

In this research, a 50 cm × 50 cm plastic board was used to contain the samples during each shooting session, after which manual counting was performed. Subsequently, image capture was conducted, as illustrated in Figure 6. All images that had been manually counted were then utilized as a training set to assess the model's accuracy in identifying real-world scenarios and to compare it with other models.

Fig. 6 - Wheat image

RESULTS

In this study, our goal is to develop an efficient wheat spike detection algorithm by improving the original YOLOv8 network. These improvements include adjusting the network architecture of YOLOv8, incorporating SPDConv and PSA attention mechanisms, and embedding them into a semi-supervised algorithm. To elucidate the impact of each improvement on the network, ablation experiments were conducted under the same training environment and hyperparameters. The YOLOv8n model served as the baseline, against which the previously mentioned improvements were sequentially implemented.

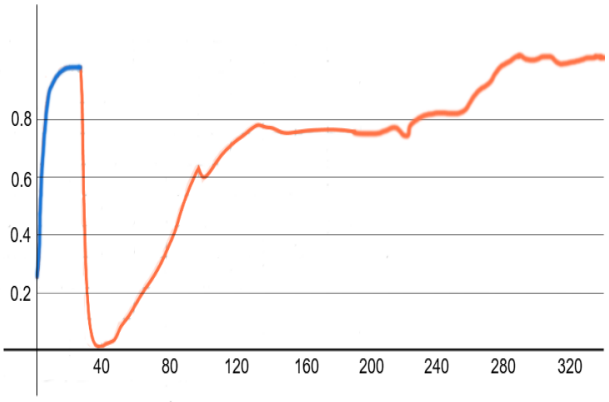

Fig. 7 - Changes in mAP50 metric during semi-supervised training stage

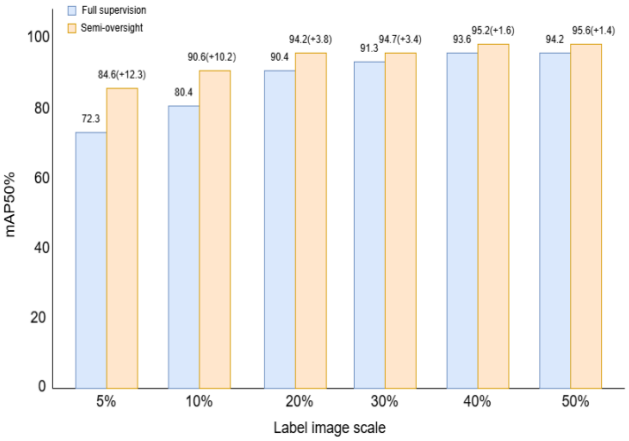

Fig. 8 - mAP50 metric of different annotation sample ratios

To verify the effectiveness of the proposed improvement strategy in improving the model accuracy, the ablation experiment was conducted based on the YOLOv8n algorithm. The experimental results are detailed in Table 1. Among them, A represents the introduced SPDConv module, and B represents the PSA attention mechanism added after SPPF, while C represents the semi-supervised object detection method mentioned in this article.

According to the data in Table 1, the detection performance is significantly improved with the gradual integration of the improved module YOLOv8n. However, the proposed algorithm did not achieve the best performance in recall rates. The reason for this phenomenon is that the algorithm improves the extraction ability of the small target features, but also optimizes the computational complexity and memory footprint. Nevertheless, a slight decrease in recall is acceptable because the algorithm's accuracy and average accuracy have peaked, given the mutual constraints between accuracy (P) and recall (R).

Table 1

Results of the ablation experiment

YOLOv8n	A	B	C	P	R	mAP50	mAP50:95
√				0.891	0.732	0.886	0.477
√	√			0.913	0.818	0.901	0.484
√		√		0.898	0.801	0.887	0.472
√			√	0.891	0.732	0.881	0.493
√	√	√		0.921	0.877	0.904	0.463
√	√		√	0.886	0.823	0.890	0.462
√		√	√	0.902	0.840	0.906	0.483
√	√	√	√	0.957	0.865	0.942	0.487

For the fully supervised training, a ratio of 0.2 and 1400 samples were used, after which 5600 images were introduced as unannotated samples to perform the semi-supervised training. Fig 7 illustrates the evolution of mAP50 values during fully supervised versus semi-supervised training. As can be seen from the figure, after the full supervised training, the mAP50 index of the algorithm has basically reached the fitting state, with the accuracy of 90.4%. Then, 3200 unlabeled samples were added for semi-supervised training. At this time, the mAP50 index of the model went through the process of first reducing and then rising, and finally achieved further improvement with an accuracy rate of 94.2%. The reason for the decrease lies in the early stage of the semi-supervised training, where the accuracy of the model is low and the quality of the pseudo-labels generated by the teacher model is not high. However, as the training continues, the quality of the pseudo-labels gradually improves, and the detection accuracy of the model is correspondingly enhanced.

Table 2

Comparative experiments on object detection algorithms

Model	Size	P	R	mAP50	mAP50:95
YOLOv5s	640	0.867	0.754	0.833	0.473
YOLOv5l	640	0.899	0.776	0.825	0.466
Faster R-CNN	640	0.898	0.846	0.910	0.486
SSD	640	0.901	0.834	0.905	0.481
YOLOv8n	640	0.891	0.732	0.881	0.493
YOLOX	640	0.902	0.859	0.921	0.528
Efficient Teacher	640	0.890	0.827	0.897	0.502
V8	640	0.904	0.852	0.914	0.494
V8ps	640	0.957	0.865	0.942	0.487

Subsequently, this study compared the effects of fully supervised and semi-supervised training methods under different annotated sample proportions, and the experimental results are shown in Fig 8. In the figure, blue area on the left represents the mAP50 values obtained from the fully supervised training with the entire annotated samples, while the orange area on the right shows the mAP50 values obtained after the further semi-supervised training. Since the total number of samples used in the training is 7000, the annotated ratio of 20% corresponds to 1400 annotated samples and 5600 unannotated samples, and so on.

To further verify the performance advantages of the algorithm proposed in this paper on the wheat spike detection dataset, comparative experiments were conducted with other mainstream lightweight object detection models, including YOLOv5s *(Jocher, 2020)*, YOLOv5l, Faster R-CNN*(Ren et al., 2017)*, SSD*(Liu et al., 2016)*, YOLOv8n, YOLOX*(Z Ge et al., 2021)* and Efficient Teacher. For fairness, their default resolutions were not adjusted. The experimental results are detailed in Table 2.

In this experiment, Efficient Teacher uses the YOLOv5s detector, while V8ps represents the semi-supervised training results with the improved YOLOv8n as the baseline model, and V8 represents the semi-supervised training results with YOLOv8n as the baseline model.

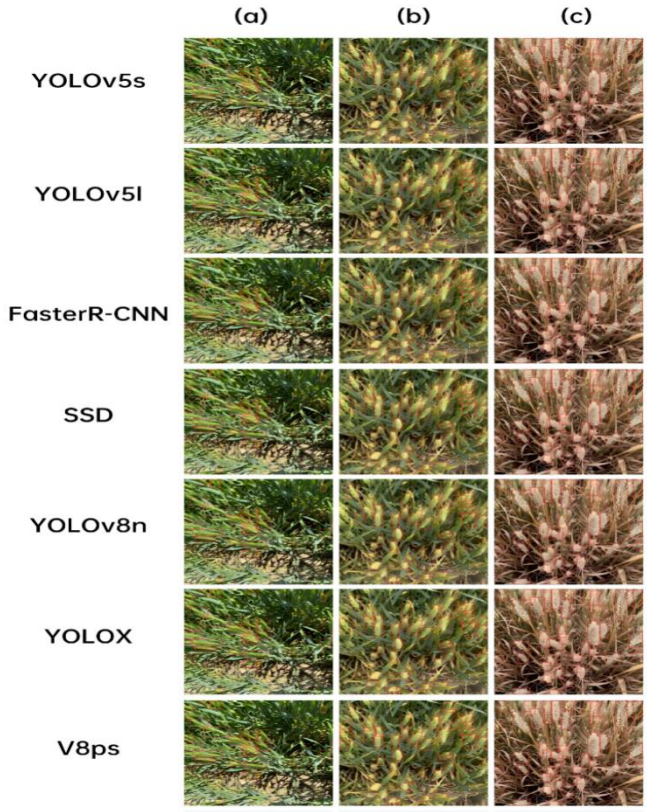

Fig. 9 - Comparison of V8ps and fully supervised models in wheat spike detection effectiveness

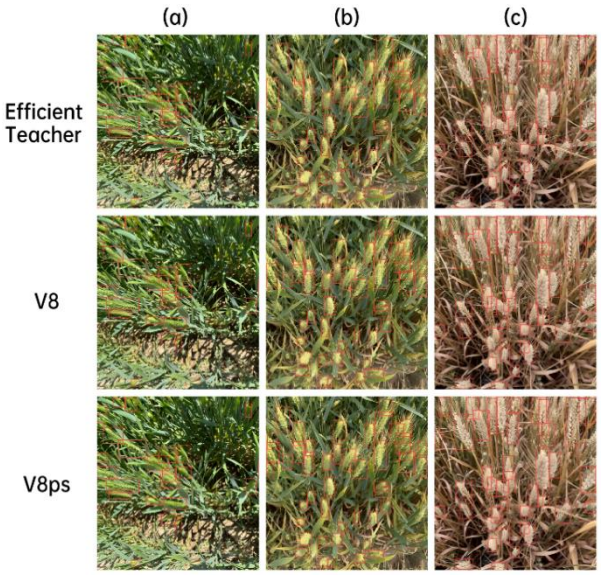

Fig. 10 - Comparison of V8ps and semi-supervised models in wheat spike detection effectiveness

The fully supervised target detection algorithms listed in the table were all obtained with 300 epochs without pre-training weights. According to the data in Table 2, using the improved YOLOv8n as the base model, the mean accuracy (mAP50) reached 94.2% under the semi-supervised training condition, and the performance improved by 4.5% compared with the semi-supervised base model Efficient Teacher. Its precision exceeds all other algorithms in the table compared to the fully supervised model. If the base model is changed from an improved YOLOv8n to an unimproved YOLOv8n, the improved YOLOv8n achieves a performance improvement of 2.8% on the mAP50 index.

Table 3

Comparative experiments on object detection algorithms

Model	Manual count	Model count	Number of missed detection	Omission factor	Precision ratio
YOLOv5s	359	321	38	10.63%	89.37%
YOLOv5l	421	379	42	10.06%	89.94%
Faster R-CNN	433	391	42	9.68%	90.32%
SSD	217	196	21	9.59%	90.41%
YOLOv8n	189	172	17	9.15%	90.85%
YOLOX	285	259	26	8.98%	91.02%
Efficient Teacher	295	269	26	8.83%	91.17%
V8	311	285	26	8.28%	91.72%
V8ps	354	331	23	6.43%	93.57%

The manually counted images were randomly categorized into nine groups, each consisting of ten images, and the model was evaluated. As shown in Table 3, the enhanced model demonstrated significantly higher accuracy in real-world situations compared to other popular models, achieving a detection accuracy of 93.57% in the tests.

Furthermore, Fig.9 and Fig.10 respectively present a comparative analysis of the wheat spike detection effects of the improved model compared to the fully supervised model and the semi-supervised model at different growth stages. In these images, (a) represents Heading stage, (b) represents Filling stage, and (c) represents Harvesting period. Accordingly, it can be concluded that the improved model in this study is more suitable for the application scenario of wheat ear detection in the field.

CONCLUSIONS

To address the challenges of small target sizes, severe occlusion, and high annotation costs in wheat spike detection, a semi-supervised wheat spike detection algorithm, named YOLOv8ps, was proposed. It is based on YOLOv8. Our main contributions are summarized as follows: wheat spike images are sourced from wheat field environments and the global wheat public database. After annotating the images, data augmentation was performed to create the final dataset. YOLOv8ps was developed by integrating the SPDConv module, which mitigates the adverse effects of processing small objects and low-resolution images, thereby enhancing the model's performance and stability in complex scenarios. Simultaneously, the introduction of the PSA attention mechanism further optimizes the model's performance. Ultimately, the combination of an improved semi-supervised learning method significantly enhances the model's detection accuracy. First, ablation experiments were conducted in a consistent training environment to analyze the independent impact of each improvement on the model. Subsequently, comparative experiments were carried out using both full-supervision and semi-supervision training methods under varying proportions of annotated samples to verify the effectiveness of the newly proposed YOLOv8ps in semi-supervised training. Finally, the model was compared with other existing models in terms of performance. Experimental results indicate that, compared to Efficient Teacher, the improved model's mean average precision (mAP50) increased by 4.5%, reaching 94.2%. In comparison to the full-supervision baseline model YOLOv8n, it increased by 6.1%, achieving excellent detection accuracy. Although model lightweighting and other optimization techniques have not yet been considered, future research will focus on expanding the dataset and exploring how to integrate these techniques with more advanced and efficient technologies to achieve a higher level of wheat detection performance.

ACKNOWLEDGEMENT

This work was supported by the Key R&D Program Project of Shanxi Province and Outstanding Innovation Project for Graduate Students in Shanxi Province under the topics are: Research on Precision Agriculture Intelligent Decision-making System [project number: 202202140601021]; and based on 5G three-dimensional intelligent cultivation system [project number:2023SJ120].

REFERENCES

[1] David, E., Serouart, M., Smith, D., Madec, S., Velumani, K., Liu, S., Wang, X., Pinto, F., Shafiee, S., Tahir, I.S.A., 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, M.A., Pozniak, C., LeBauer, D.S., Lillemo, M., Poland, J., Chapman, S., de Solan, B., Baret, F., Stavness, I., Guo, W., (2021). Global Wheat Head Detection 2021: An Improved Dataset for Benchmarking Wheat Head Detection Methods. *Plant Phenomics 2021*. https://doi.org/10.34133/2021/9846158

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

[3] Fernandez-Gallego, J.A., Kefauver, S.C., Gutiérrez, N.A., Nieto-Taladriz, M.T., Araus, J.L., (2018). Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. *Plant Methods*. 14, 22. https://doi.org/10.1186/s13007-018-0289-4

[4] Ge Z.， Liu S.， Wang F.， Li Z.， Sun J., (2021). *YOLOX: Exceeding YOLO Series* in 2021.

[5] Jingbo F., (2021), *Phenotypic Traits Extraction and Analysis of Potted Wheat Based on Deep Learning*. *（基于深度学习的盆栽小麦表型性状提取与分析）*, PhD dissertation, Huazhong Agricultural University, Wuhan/China.

[6] Jocher, Glenn R., Alex Stoken, Jiří Borovec, NanoCode, Ayushi Chaurasia, TaoXie, Changyu Liu, Abhiram, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang, Jan Hájek, Laurentiu Diaconu, Marc, Yonghye Kwon, Oleg, wanghaoyang, Yann Defretin, Aditya Lohia, ml ah, Ben

Milanko, Ben Fineran, D. P. Khromov, Ding Yiwei, Doug, Durgesh and Francisco Ingham. Ultralytics yolov5: v5.0 - YOLOv5 - P6 1280 models. *Zenodo* (2020). https://doi.org/10.5281/zenodo.4154370

[7] Li, Y., Du, S., Yao, M., Yi, Y., Yang, J., Ding, Q., He, R., (2018). Method for wheatear counting and yield predicting based on image of wheatear population in field. （基于小麦群体图像的田间麦穗计数及产量预测方法）*Transactions of the Chinese Society of Agricultural Engineering* 34, 185–194.

[8] Liu Tao, Sun Chengming, Wang Lijian., (2014). Wheat spike counting based on image processing technology. （基于图像处理技术的大田麦穗计数）*Transactions of the Chinese Society for Agricultural Machinery*, 2014, 45(2): 282-290.

[9] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., (2016). SSD: Single Shot MultiBox Detector. *European Conference on Computer Vision* (2015). pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[10] Liu, Z., Huang, W., Wang, L., (2019). Field wheat ear counting automatically based on improved K-means clustering algorithm. （基于改进 K-means 聚类算法的大田麦穗自动计数）*Transactions of the Chinese Society of Agricultural Engineering* 35, 174–181. https://doi.org/10.11975/j.issn.1002-6819.2019.03.022

[11] Lyu, J., Li, S., Zeng, M., Dong, B., (2022). Detecting bagged citrus using a semi-supervised SPM-YOLOv5. （基于半监督 SPM-YOLOv5 的套袋柑橘检测算法）*Transactions of the Chinese Society of Agricultural Engineering.* 38, 204–211. https://doi.org/10.11975/j.issn.1002-6819.2022.18.022

[12] Ren, S., He, K., Girshick, R., Sun, J., (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* 39, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[13] Sunkara, Raja and Tie Luo. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. (2022) *ArXiv* abs 2208.03641: pp. 443–459. https://doi.org/10.1007/978-3-031-26409-2_27

[14] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G., (2024). YOLOv10: Real-Time End-to-End Object Detection. (2024). *ArXiv* abs 2405.14458. https://doi.org/10.48550/arXiv.2405.14458

[15] Xu, B., Chen, M., Guan, W., Hu, L., (2023). Efficient Teacher: Semi-Supervised Object Detection for YOLOv5.*ArXiv* abs 2302.07577. https://doi.org/10.48550/arXiv.2302.07577

[16] Yang, G., Wang, J., Nie, Z., Yang, H., Yu, S., (2023). A Lightweight YOLOv8 Tomato Detection Algorithm Combining Feature Enhancement and Attention. *Agronomy* 13, 1824. https://doi.org/10.3390/agronomy13071824

[17] Zhou, H., Jiang, F., Lu, H., (2023). SSDA-YOLO: Semi-supervised domain adaptive YOLO for cross-domain object detection.*Computer Vision and Image Understanding* 229, 103649. https://doi.org/10.1016/j.cviu.2023.103649