MILLET EAR DETECTION METHOD IN UAV IMAGES BASED ON IMPROVED YOLOX
/
基于改进 YOLOX 的无人机图像谷穗检测方法

Fuming MA[1)], Shaonian LI*[1)], Juxia LI[2)], Yanwen LI[2)], Lei DUAN[2)], Linwei LI [2)], Jing TAN [1)], Yifan WANG [1)]
[1)] College of Energy and Power Engineering, Lanzhou University of Technology, Lanzhou, Gansu/ China
[2)] College of Information Science and Engineering, Shanxi Agricultural University, Jinzhong, Shanxi/ China
Tel: 0931-2973750; E-mail: Lsn19@163.com
Corresponding author: Shaonian Li
DOI: https://doi.org/10.35633/inmateh-75-59

Keywords: *YOLOX; CBAM; EIoU; millet ear detection; UAV; deep learning*

ABSTRACT
Rapid and accurate detection of millet ears is essential for yield estimation and phenotypic studies. However, traditional detection methods primarily rely on manual observation, which are both subjective and labor-intensive. To address this issue, this study employed Unmanned Aerial Vehicle (UAV) for image data collection of millet ears and proposed the YOLOX-CBAM-EIoU model to facilitate real-time detection, focusing on challenges such as small millet ears size, dense distribution, and severe occlusion in the dataset. Firstly, the Mosaic data augmentation technique was employed to enhance the diversity of the dataset. Subsequently, the CBAM attention mechanism was incorporated between the Neck and Prediction layers of YOLOX, enabling the reallocation of channel weights to enhance the extraction of fine-grained features and deeper semantic information. Additionally, EIoU Loss was utilized as the loss function for bounding box regression to mitigate missed detections in dense scenes. The improved model achieved an average precision (AP) of 90.30%, a 6.44 percentage point increase over the original YOLOX model, significantly enhancing detection performance for densely distributed millet ears. The improved model also demonstrated a Precision of 91.01%, Recall of 89.45%, and F1-score of 90.22, highlighting strong robustness and generalization capabilities. These findings substantiate the efficacy of the YOLOX-CBAM-EIoU model in improving detection performance under dense distribution and occlusion conditions, providing valuable technical reference for further UAV-based analyses of millet ears phenotypes and yield predictions.

摘要
小米穗的快速准确检测对于产量估计和表型研究至关重要。然而，传统的谷穗检测主要依靠人工观察，不仅主观性强且耗时耗力。为此，本研究通过无人机进行谷穗图像数据采集，主要针对数据集中谷穗体积小、分布密集、遮挡严重等问题提出了 YOLOX-CBAM-EIoU 模型对谷穗进行实时检测。该模型首先在 YOLOX 的颈部层和预测层之间引入了 CBAM 注意力模块，通过重新分配不同通道的权重，获得了更浅层的细粒度特征和更深层的语义信息，以提高对谷穗表型的特征提取能力；其次，采用 EIOU 函数作为回归损失函数，以改善密集场景下谷穗目标的漏检问题。结果表明，改进后的模型检测平均精度（AP）达到 90.30%，与原 YOLOX 模型相比提高了 6.44 个百分点，显著提高了密集分布的谷穗目标检测性能。改进后模型的精确率达到 91.01%，召回率达到 89.45%，F1 分数达到 90.22，表现出较强的鲁棒性和泛化能力。结果充分证明了 YOLOX-CBAM-EIoU 模型能显著提高谷穗在密集分布及遮挡条件下的检测效果，为进一步使用无人机分析谷穗表型和产量预测提供了技术参考。

INTRODUCTION
Millets are highly drought-resistant and can thrive in poor soils, making it an important coarse grain crop globally. China's millet planting area accounts for about 80% of the world, and its production accounts for about 90% of the world's total output *(Li et al., 2021)*. Millet ears are crucial agronomic indicators for assessing yield and quality, playing a key role in breeding, nutritional diagnosis, and growth period monitoring. Therefore, research and management of millet ears could improve millets yield and quality, providing scientific basis for agricultural production.

[1] *Fumin Ma, M.S. Stud. Eng.; Shaonian Li, Prof. Ph.D. Eng.; Juxia Li, Prof. Ph.D. Eng.; Yanwen Li, Lecturer M.S. Eng.; Lei Duan, M.S. Stud. Eng.; Linwei Li, Lecturer M.S. Eng.; Jing Tan, M.S. Stud. Eng.; Yifan Wang, B.S. Stud. Eng.*

The traditional method for detecting millet ears primarily relies on manual observation, which is labor-intensive, subjective and inefficient. In recent years, with the advancement of computer technology, the integration of computer vision technology with agricultural production has deepened, demonstrating significant application potential in the field of crop detection. *Zhao et al., (2014),* proposed a method for wheat ear detection that integrated color information with the AdaBoost algorithm, achieving automated recognition and counting of wheat ears. *Liu et al., (2014*), proposed a wheat ear counting method based on image analysis technology. By integrating color and texture features for image segmentation, the method achieved counting accuracies of 95.77% and 96.89% under broadcast and row-seeding conditions, respectively. *Li, (2016)* applied binarization and morphological processing to raw wheat images, incorporating the Harris corner detection algorithm to extract the wheat ear skeleton. This method effectively enabled the automatic assessment of wheat ear density per unit area. *Li et al., (2018),* converted RGB images of wheat ears to binary images through color space transformation. They then applied boundary and regional feature parameters to identify fused regions. Subsequently, a line-matching technique based on concavity point detection was employed to segment these fused regions, enabling accurate quantification of the wheat ears. *Meng et al., (2019),* utilized the improved K-means algorithm to cluster wheat texture features and achieve the recognition of wheat ears. When the traditional image processing methods shown above were used for target detection of cereal crops, it was found that the traditional methods had strong dependence on texture and color, making them highly susceptible to variations in background, lighting, and foliage. Additionally, these methods often required manual adjustment of feature thresholds, which could introduce subjective bias and lead to poor generalization across diverse environments and conditions.

Given the limitations of traditional machine learning technology, crop object detection methods based on deep learning technology have gradually emerged as the mainstream solution. Current research primarily concentrates on the application of deep learning to wheat, rice, and other crops, focusing on improving the accuracy and efficiency of detection models (*Yang et al., 2022; Huang et al., 2022; Liu et al., 2023; Xiong, 2018; Yang et al., 2021*). *Duan et al., (2018)*, proposed a rice panicle segmentation network model based on SegNet, which effectively addressed the challenges to segmentation accuracy posed by the irregular panicle edges, substantial variations in appearance across different rice varieties and growth stages, and occlusions. Zhang et al. proposed a convolutional neural network model for the identification of winter wheat ear. By incorporating non-maximum suppression technology, this model achieved the rapid and accurate detection of wheat ear in field conditions (*Zhang et al., 2019*). Bao et al. proposed a wheat ear recognition model based on convolutional neural network. This model incorporated an image pyramid to construct multi-scale sliding windows and utilized non-maximum suppression techniques to eliminate overlapping bounding boxes, ultimately achieving efficient counting of wheat ears (*Bao et al., 2019*). *Zhang et al.,* (*2021),* proposed an improved wheat ear detection method based on the feature pyramid network, which weighed the underlying high-resolution feature map to enhance the useful information channel. *Zhang et al., (2021),* proposed a rice panicle detection model based on Faster R-CNN, which enhanced the model's performance in detecting small objects by incorporating dilated convolution. In 2023, *Bao et al.* proposed a wheat ear detection model based on TPH-YOLO, which effectively localized the wheat ear in high-density environments (*Bao et al., 2023*). In the same year, *Cai et al. (2023)*, made adaptive improvements to the YOLOv5l model to enable precise detection of rice panicles in field environments. By incorporating the Efficient Channel Attention mechanism before the spatial pyramid pooling layer of the original YOLOv5l model, this study effectively enhanced the model's accuracy and speed in detecting small targets. *Cai et al., (2023),* proposed a method to reduce the training data for sorghum panicle detection through semi-supervised learning. The results indicated that this approach achieved performance comparable to supervised methods while using only 10% of the original training data.

However, the application of deep learning techniques in crop detection has primarily focused on wheat, rice, and other crops, while research on millet ear detection remained limited. On the other hand, the small size, dense distribution, and frequent occlusions of millet ears in field environments present significant challenges to accurate detection. To address these issues, this study employed UAV to capture images of millet ears in field environments and constructed a dataset of millet ears. Additionally, a millet ear detection model based on YOLOX was proposed. Specifically, (1) the CBAM attention mechanism was introduced to improve the YOLOX model’s ability to extract detailed features of millet ears; (2) the original IoU loss function w replaced with the EIoU loss function, further improving the model's performance in detecting small, densely distributed millet ear targets.

MATERIALS AND METHODS

Data acquisition and processing

Data acquisition

The millet images used in this study were acquired from an experimental field located on the southern side of Shanxi Agricultural University, Jinzhong City, Shanxi Province (37°42′45″ N, 112° 59′ 27″ E). The millet cultivar employed was Jingu21. Data collection was conducted between July and August 2022, during the heading stage of the millet. A DJI MAVIC AIR 2 portable UAV, equipped with a 48-megapixel visible light camera, was utilized for image acquisition. The camera was oriented vertically, at a 90° angle to the ground. The UAV was flown at an altitude of 3 m and a speed of 5 m/s during the image capture process. Representative images of the millet are presented in Fig. 1.

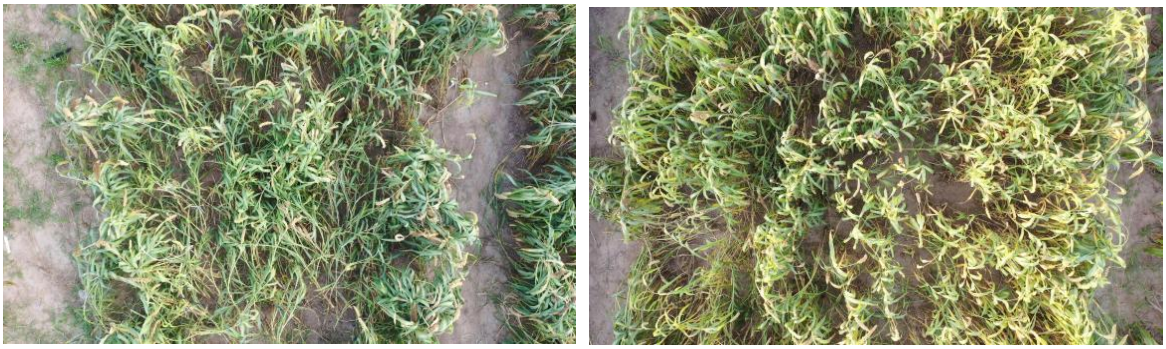


Fig. 1 - Images collected by UAV

Data Preconditioning

To ensure the high quality of the dataset, video footage was recorded under optimal conditions, specifically during clear weather and adequate lighting, between 9:30 AM and 10:30 AM. The raw videos were processed through frame extraction, with the elimination of images that were either highly similar or blurry, which culminated in the selection of 117 images, each with dimensions of 3840 pixels by 2160 pixels. However, the pixel proportion occupied by millet ears in these original images was found to be insufficient, presenting challenges for effective model training. Additionally, processing the original images at this resolution resulted in excessively long times. To address these concerns, each original image was subdivided into six smaller images using a 3×2 grid, yielding a total of 702 images, each with a resolution of 1280 pixels by 1080 pixels. An illustration of this image segmentation process is provided in Fig. 2.

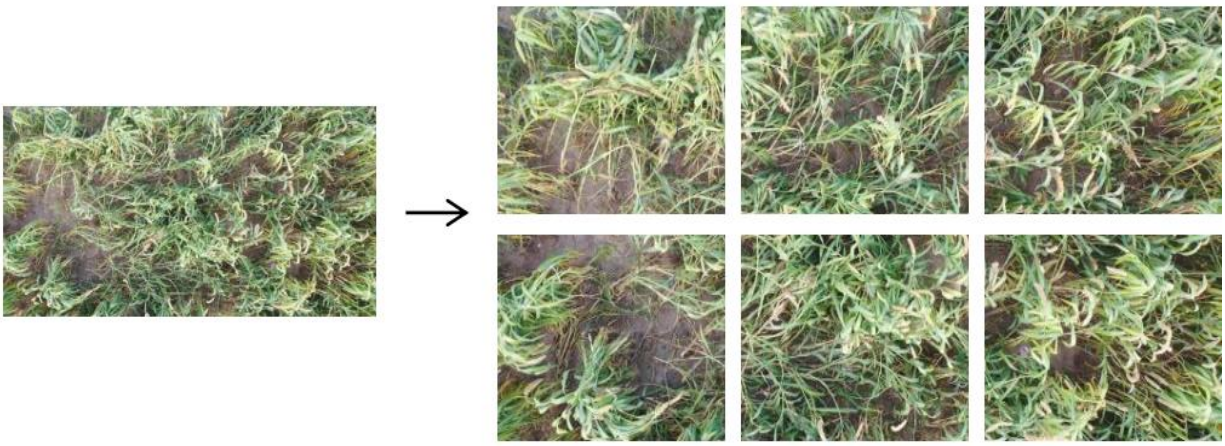


Fig. 2 - Image cutting example diagram

The millet ears in the images were annotated using Labeling software, and the annotation result is shown in Fig. 3.

Fig. 3 - Data annotation results

In this study, the Mosaic data augmentation technique was employed to enhance the diversity of the dataset. The Mosaic data augmentation method involved a series of operations, including random scaling, rotation, and stitching. The newly generated image was created by stitching four original images after applying the transformations, as illustrated in Fig. 4. A total of 2,106 images of millet ears were obtained and divided into training, validation, and test sets in an 8:1:1 ratio.

Fig. 4 - Mosaic data enhancement

Algorithm introduction

YOLOX-CBAM-EIoU network architecture

In this study, an improved model named YOLOX-CBAM-EIOU was proposed for millet ear detection. Firstly, a Convolutional Block Attention Module (CBAM) was integrated between the Neck and Prediction layers of the YOLOX model. Secondly, the EIoU function was introduced as the model's loss function. These modifications were found to effectively improve the stability of the bounding box regression and mitigate the scattering of IoU loss during training, thereby improving the overall detection performance. The network architecture of the improved model is illustrated in Fig. 5.

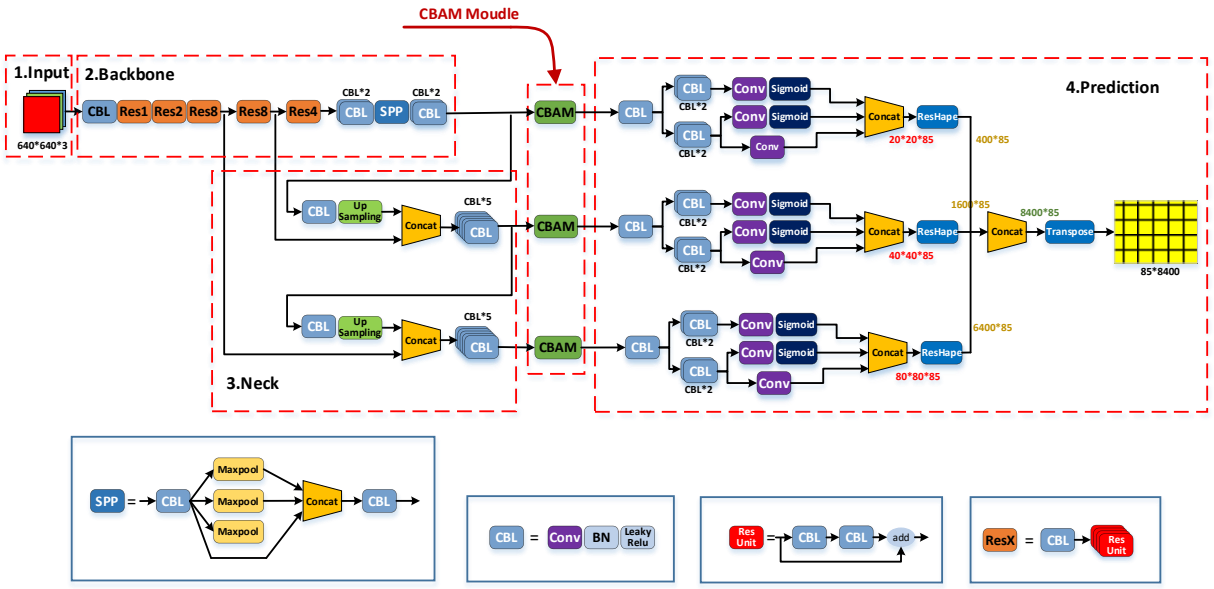


Fig. 5 - YOLOX-CBAM-EIoU network architecture

YOLOX network architecture

The YOLOX network architecture is mainly composed of four parts: the Input, the Backbone, the Neck, and the Prediction (*Ge et al., 2021*). In the Input stage, the image undergoes scaling, enhancement, and normalization processes to prepare it for further analysis. The Backbone utilizes CSPDarknet53 (Cross-Stage Partial Darknet53) for feature extraction, incorporating Spatial Pyramid Pooling (SPP) to effectively capture multi-scale feature information. The Neck component incorporates the Path Aggregation Network (PANet) structure, which enhances the generation of high-quality feature maps crucial for subsequent prediction tasks. In the Prediction stage, the network comprises three decoupled heads, each characterized by a unique configuration of convolutional layers and activation functions. This architecture provides a robust and efficient solution for object detection and localization across diverse scales and contexts.

CBAM attention mechanism

The CBAM is composed of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM) *(Woo et al., 2018)*. As illustrated in Fig. 6, the architecture of CBAM integrates two modules to enhance the network's representational power. Traditional convolutional neural network-based attention mechanisms primarily focus on channel-wise interactions, concentrating on the analysis of channel dimensions. In contrast, CBAM incorporates attention mechanisms in both channel and spatial dimensions, thus implementing a sequential attention structure that progresses from channel to spatial dimensions. The spatial attention mechanism enables the network to concentrate on pixel regions critical for image classification while disregarding less important areas. Simultaneously, the channel attention mechanism manages the distribution of channels within the feature maps. The synergistic combination of these two attention dimensions significantly improves the model's performance.

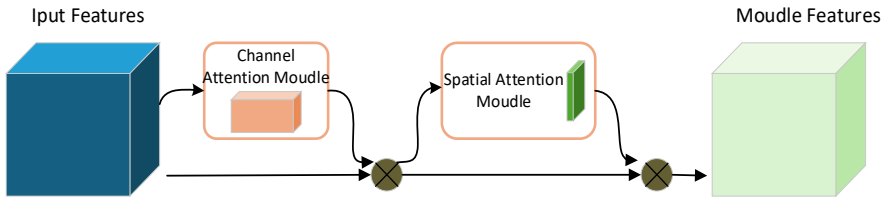


Fig. 6 - Structure of CBAM attention mechanism

EIoU loss function

In this study, the EIoU loss function *(Zhang et al., 2022)* was employed to replace the original IoU loss function in the YOLOX model. The definition of EIoU loss is shown in Equation (1):

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp} = 1 - IoU + \frac{\rho^2(\mathbf{b},\mathbf{b}^{gt})^2}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w,w^{gt})^2}{(w^c)^2} + \frac{\rho(h,h^{gt})}{(h^c)^2} \quad (1)$$

where:

$\mathbf{b}$ and $\mathbf{b}^{gt}$ denote the central points of anchor box and target box; $\rho^2(\cdot) = ||\mathbf{b} - \mathbf{b}^{gt}||$ indicates the Euclidean distance; w and w^{gt} denote the weight of anchor box and target box; h and h^{gt} denote the height of anchor box and target box; w^c and h^c are the width and height of the smallest enclosing box covering the two boxes.

Evaluation indicators

In this study, the evaluation indicators for the model included Precision (P), Recall (R), F1-score, and Average Precision (AP). The formulas for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (4)$$

$$AP = \int_0^1 P(R)dR \quad (5)$$

where:

P denotes accuracy; R denotes recall; $F1$ denotes $F1$-score; AP denotes average precision; TP indicates the number of instances where the model correctly identifies the grain bounding box, matching the actual category label; FP denotes the number of instances where the model incorrectly identifies the grain bounding box, resulting in a mismatch with the actual category label; FN represents the instances where the model fails to detect any millet ears.

Parameter settings

The experimental equipment and configuration are shown in Table 1.

Table 1

Experimental Configuration

Parameter	Configuration
Operating System	Windows 10
CPU	11th Gen Intel(R) Core (TM) i7-11800 H
GPU	RTX 3080 Ti
Memory	16 GB
Programming Language	Python 3.8
Development Environment	PyCharm
CUDA Driver	11.7

In the comparative experiments conducted in this study, all detection models were configured with identical hyperparameters and trained on the same dataset. Specifically, the training was conducted for 200 epochs, with a momentum of 0.9 and a weight decay regularization coefficient of 0.005. A learning rate decay strategy was adopted, with an initial learning rate set to 0.0001.

RESULTS AND ANALYSIS

Comparison of detection performance of different models

In this section, three comparative experiments were designed and conducted under identical experimental conditions to identify the millet ear detection model that yields the best performance. The details of the experimental procedures and the resulting analyses were presented below.

In the first set of comparative experiments, classical detection models including EfficientDet (*Tan et al., 2020*), YOLOv4 (*Bochkovskiy et al., 2020*), YOLOv5, and YOLOX were selected for millet ear detection. The detection results are presented in Table 2.

Table 2

Comparative experimental results of different models

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)
EfficientDet	78.11	52.04	62.46	70.50
YOLOv4	82.18	66.90	73.76	81.90
YOLOv5	83.13	62.73	71.50	81.76
YOLOX	88.07	54.95	67.68	83.86

Based on the comparison results presented in Table 2, it could be seen that the YOLOX model achieved an AP of 83.86%. Compared to the EfficientDet, YOLOv4, and YOLOv5 models, YOLOX demonstrated improvements in AP of 13.36%, 1.96%, and 2.1%, respectively. The results demonstrated the superior performance of the YOLOX model in millet ear detection tasks.

In the second set of comparative experiments, attention mechanisms including SE, BAM, NAM, and CBAM were selected to enhance the YOLOX model. These modified models were subsequently compared with the original YOLOX model to evaluate their performance improvements. The detection results are presented in Table 3.

Table 3

Results of attentional mechanism comparison experiments

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)
YOLOX	88.07	54.95	67.68	83.86

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)
YOLOX-SE	81.59	79.95	80.76	85.66
YOLOX-BAM	83.79	76.24	79.84	84.63
YOLOX-NAM	84.75	75.38	79.79	85.37
YOLOX-CBAM	82.09	79.12	80.58	86.24

Based on the comparison results presented in Table 3, the introduction of CBAM significantly enhanced model performance compared to other attention mechanisms. Specifically, the YOLOX-CBAM model achieved a 12.9% improvement in the F1 and a 2.38% increase in AP relative to the baseline YOLOX model. Furthermore, in comparison to YOLOX models that incorporated SE, BAM, and NAM modules, the YOLOX-CBAM model demonstrated improvements in AP of 0.58%, 1.61%, and 0.87%, respectively. The superior performance of CBAM could be attributed to its dual-channel and spatial attention architecture, which enriched feature representations across both dimensions by integrating max pooling and average pooling. This design significantly enhanced the model's capacity to capture critical details in millet ear images. The results indicated that the integration of the CBAM attention mechanism into the YOLOX framework facilitated more accurate and reliable detection of millet ears.

In the third set of comparative experiments, the loss functions including GIoU, CioU, DIoU and EIoU were employed to replace the IoU loss function in YOLOX. The detection results are presented in Table 4.

Table 4

Loss function comparison experiment results

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)
YOLOX	88.07	54.95	67.68	83.86
YOLOX-GIoU	77.94	42.67	55.15	64.03
YOLOX-CIoU	79.93	49.52	61.15	69.43
YOLOX-DIoU	70.97	47.89	57.19	62.16
YOLOX-EIoU	92.04	56.56	70.06	86.06

Based on the comparison results presented in Table 4, the experiments demonstrated that the YOLOX-EIoU model exhibited superior performance over all other models, as evidenced by both AP and F1. Specifically, the F1 of the YOLOX-EIoU model surpassed that of the YOLOX, YOLOX-GIoU, YOLOX-CIoU, and YOLOX-DIoU models by 2.38%, 14.91%, 8.91%, and 12.87%, respectively. Additionally, the AP of the YOLOX-EIoU model surpassed that of the YOLOX, YOLOX-GIoU, YOLOX-CIoU, and YOLOX-DIoU models by 2.2%, 22.03%, 16.63%, and 23.9%, respectively. The results indicated that the EIoU loss function serves as the optimal alternative to the IoU loss function in the original YOLOX model.

Ablative test performance comparison

To validate the improvement method proposed in this study, ablation experiments were conducted focusing on the CBAM and the EIoU loss to assess the effectiveness of each enhancement. CBAM and EIoU were sequentially integrated into the original YOLOX model while employing the same parameter configuration throughout the training process. The experimental results are presented in Table 5. The introduction of CBAM resulted in significant performance enhancement, with a 2.38% increase in AP. Furthermore, the utilization of both CBAM and EIoU led to improvements in F1 and Recall, culminating in a 6.44% increase in AP. It could be concluded that the incorporation of the attention mechanism CBAM enabled the model to selectively emphasize informative features, thereby enhancing its representational capacity, as evidenced by the marked improvement in detection accuracy. Additionally, the EIoU loss function enhanced sensitivity to the position and size of bounding boxes by introducing an extra penalty term, which further contributed to the improvement in detection precision.

Table 5

Results of ablation experiment

Model	Precision (%)	Recall (%)	F1 (%)	AP (%)
YOLOX	88.07	64.95	74.76	83.86
YOLOX-CBAM	82.09	79.12	80.58	86.24
YOLOX-EIoU	92.04	56.56	70.06	86.06
YOLOX-CBAM-EIoU	91.01	89.45	90.22	90.30

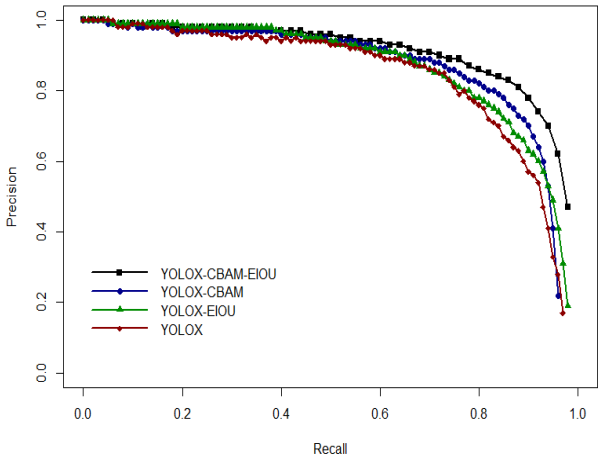
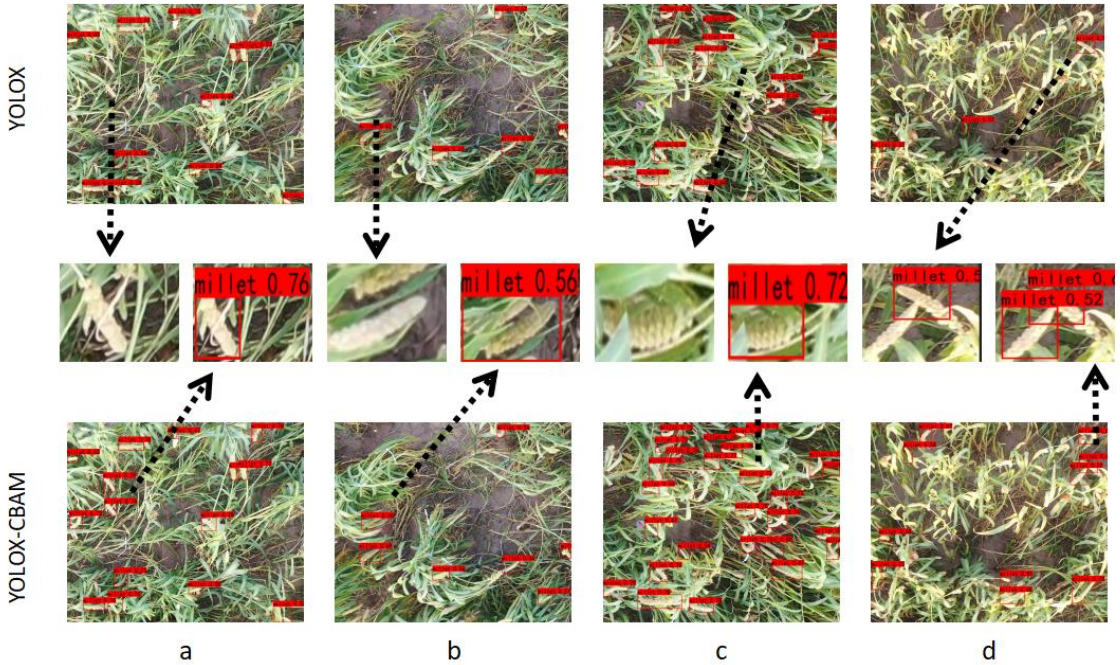


Fig. 7 - P-R curve of ablation experiment

To provide a more intuitive demonstration of the performance of the proposed YOLOX-CBAM-EIoU model, the P-R curves for the four different models mentioned in the ablation experiments were plotted, as shown in Fig. 7. In comparison to the other three models, the P-R curve of the YOLO-CBAM-EIoU model consistently lies at the top, and the area under the curve is the largest. The result indicates that the model achieves the highest AP, demonstrating superior performance in the task of millet ear detection.

Visualization analysis of detection results

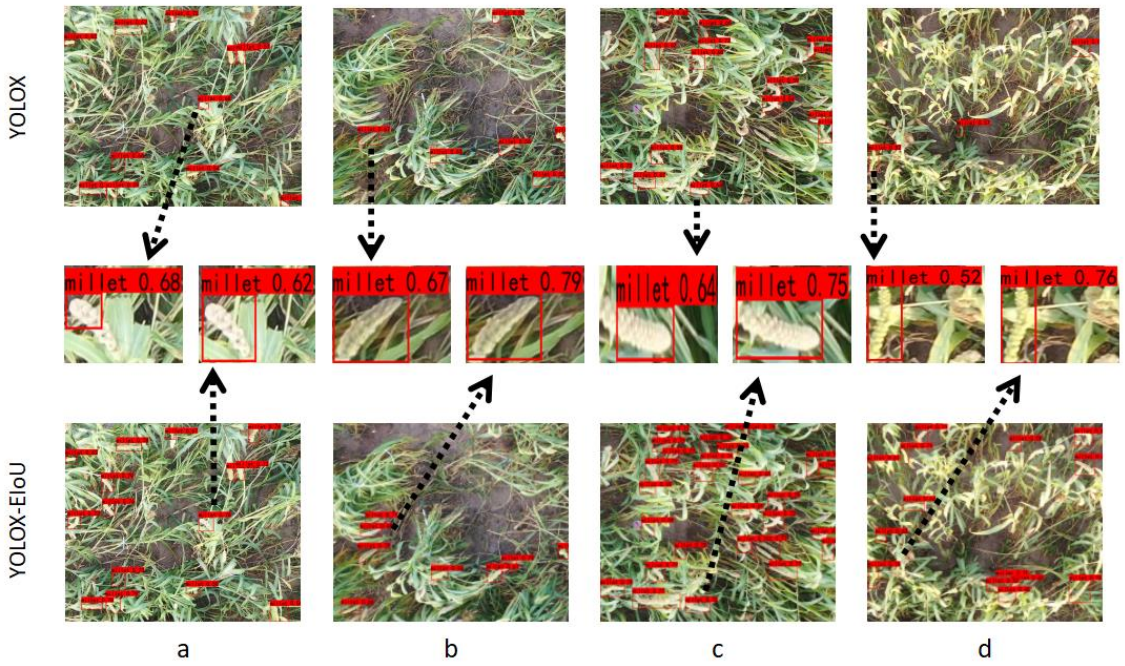
To provide a more intuitive evaluation of model performance, YOLOX, YOLOX-CBAM, YOLOX-EIoU, and YOLOX-CBAM-EIoU were tested and visually analyzed under identical experimental conditions. In the visualization results shown below, Fig. 8a and 8b represent cases where the millet ears are relatively dispersed, Fig. 8c and 8d correspond to cases where the millet ears are relatively dense and occluded. The red boxes in the figure indicate the detected millet ears.



(*Note: the upper layer is YOLOX visualization result diagram; The middle layer is the enlarged image of the contrast effect; The lower layer is the YOLO-CBAM visual result chart*)

Fig. 8 - Comparison of the detection performance before and after using the CBAM module

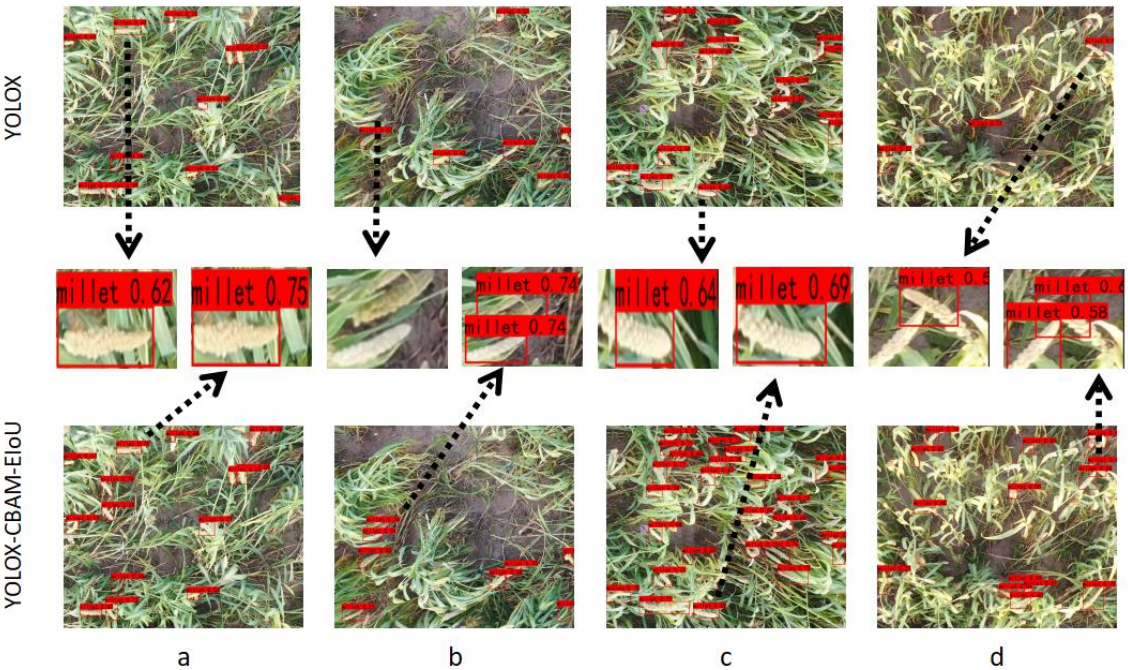
The detection results of the YOLOX-CBAM model are shown in Fig. 8. It can be observed that the millet ears in UAV images are small and subject to occlusion both among the ears and between ears and leaves. This causes considerably missed detection with the original YOLOX model. The incorporation of the CBAM attention mechanism can significantly reduce these missed detections, as evidenced in Fig. 8a and 8b. Moreover, the CBAM attention mechanism can effectively mitigate the issue of missed detection resulting from occlusion between millet leaves and ears, as shown in Fig. 8c and 8d.



(*Note: the upper layer is the YOLOX visualization result diagram; The middle layer is the enlarged image of the contrast effect; Below is the YOLO-EIoU visualization result chart*)

Fig. 9 - Comparison of detection performance before and after using the EIOU loss function

The detection results of the YOLOX-EIoU model are shown in Fig. 9. It can be observed that the YOLOX model exhibits instances of missing detections, along with relatively low confidence scores. Conversely, the improved YOLOX-EIoU model has fewer occurrences of missed detections, and a notable enhancement in the confidence of the predicted bounding boxes.



(*Note: the upper layer is YOLOX visual result diagram; The middle layer is the enlarged image of the contrast effect; The lower layer is the YOLO-CBAM-EIoU visual result chart*)

Fig. 10 - Comparison of detection performance before and after combining CBAM with EIoU

The detection results of the YOLOX-CBAM-EIoU model are shown in Fig. 10. It can be observed that our improved model can significantly reduce the occurrence of false negatives when detecting small-sized millet ear targets and partially occluded millet ear targets.

Furthermore, the discrepancy between the predicted bounding boxes and the actual bounding boxes is markedly minimized, with confidence levels achieving the highest standards among all tested models.

CONCLUSIONS

To address the current challenges in millet ear detection, such as the small size of millet ears, their dense distribution, and severe occlusion, this study proposed an improved model based on the YOLOX model. The proposed model achieved improvements of 2.94% in precision, 24.5% in recall, 15.46% in F1-score, and 6.44% in average precision. The following improvements were made in this study:

(1) CBAM attention mechanism was added to the neck of the YOLOX model to enhance its ability to extract detailed features of millet ears.

(2) The IoU loss function was replaced with the EIoU loss function. This refinement enhanced the precision of box dimensions and positions, leading to improved detection accuracy and reduced issues with low confidence and misalignment between ground truth and predicted boxes.

(3) The YOLOX-CBAM-EIoU model was proposed. Experiments demonstrated that integrating the CBAM attention mechanism with the EIoU loss function achieved the highest detection performance. The proposed improvement method could address issues observed in real-world application environments, including missed detections, false positives, significant discrepancies between ground truth and predicted boxes, as well as low confidence in detection boxes.

The experiment results demonstrate that this model could effectively solve the problem and accurately detect millet ears in the UAV image in the field environment, which provides a way for realizing intelligent management of large-scale millet planting. Furthermore, the model would be used in the planting and management of more cereal crops, and further research would be conducted on the calculation of crop density to provide a reference for the construction of intelligent farms.

ACKNOWLEDGEMENTS

This project is financially supported by the National Nature Science Foundation of China (52165006), Fundamental Research Program of Shanxi Province (202203021212450). The authors declare no competing interests.

REFERENCES

[1] Bao L., Wang M., Liu J., Wen B., Ming Y., (2019), Estimation Method of Wheat Yield Based on Convolution Neural Network (基于卷积神经网络的小麦产量预估方法), *Acta Agriculturae Zhejiangensis*, 32(12), 2244-2252.

[2] Bao W., Xie W., Hu G., Yang X., Su B., (2023), Wheat Ear Counting Method in UAV Images Based on TPH-YOLO (基于 TPH-YOLO 的无人机图像麦穗计数方法), *Transactions of the Chinese Society of Agricultural Engineering*, 39(01), 155-161.

[3] Bochkovskiy A., Wang, C., Liao, H., (2020), Yolov4: Optimal Speed and Accuracy of Object Detection, *arXiv preprint arXiv: 2004. 10934.*

[4] Cai, E., Guo, J., Yang, C., Delp, E., (2023), Semi-Supervised Object Detection for Sorghum Panicles in UAV Imagery, *In IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, CA/USA, pp. 6482-6485.

[5] Cai Z., Cai Y., Zeng F., Yue, X., (2023), Rice Panicle Recognition in Field Based on Improved YOLOv5l Model (基于改进 YOLOv5l 的田间水稻稻穗识别), *Journal of South China Agricultural University*, 45(01), 108-115.

[6] Duan L., Xiong X., Liu Q., Yang W., Huang, C., (2018), Field Rice Panicle Segmentation Based on Deep Full Convolutional Neural Network (基于深度全卷积神经网络大田稻穗分割), *Transactions of the Chinese Society of Agricultura Engineering*, 34(12), 202-209.

[7] Ge Z., Liu S., Wang F., Li Z., Sun J., (2021), YOLOX: Exceeding YOLO Series in 2021, *arXiv preprint arXiv: 2107. 08430.*

[8] Han J., Yuan X., Wang Zhun., Chen Y, (2023), UAV Dense Small Target Detection Algorithm Based on YOLOv5s (基于 YOLOv5s 的无人机密集小目标检测算法), *Journal of Zhejiang University (Engineering and Technology)*, 57(06), 1224-1233.

[9] Huang Z., (2022), *Research on Wheat Yield Estimation Based on Deep Learning Ear Recognition* (基于深度学习麦穗识别的小麦估产研究), MSc dissertation, Shandong Agricultural University, Shandong/China.

[10] Li S., Liu F., Liu M., Cheng R., Xia E., Diao X., (2021), Current Status and Future Prospective of Foxtail Millet Production and Seed Industry in China (中国谷子产业和种业发展现状与未来展望), *Agricultural Sciences in China Scientia*, 54(03), 459-470.

[11] Liu S., Ge H., Duan J., (2023), YOLOv5 Wheat Detection Algorithm with Lightweight Convolution and Attention Mechanism (结合轻量卷积和注意机制的 YOLOv5 麦穗检测算法), *Modern Computer*, 29(09), 32-38.

[12] Liu T., Sun C., Wang L., Sun Z., Zhu X., Guo W. (2014). In-field Wheat ear Counting Based on Image Processing Technology (基于图像处理技术的大田麦穗计数), *Transaction of the Chinese Society for Agricultural Machinery*, 45(02), 282-290.

[13] Li Y., Du S., Yao M., Yi Y., Yang J., Ding Q., et al., (2018), Method for Wheat Ear Counting and Yield Predicting Based on Image of Wheat Ear 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(21), 185-194.

[14] Li Z., (2016), *Research on Wheat Spike Identification Based on Color Features (基于颜色特征的麦穗图像识别算法研究)*, MSc dissertation, Henan Agricultural University, Henan/China.

[15] Meng L., (2019), *Research on Field Wheat Ear Recognition Technology Based on Color and Texture Characteristics (基于颜色和纹理特征的大田麦穗识别技术研究)*, MSc dissertation, Henan Agricultural University, Henan/China.

[16] Tan M., Pang R., Le Q.V., (2020), EfficientDet: Scalable and Efficient Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle/United States, pp. 10781-10790.

[17] Woo S., Park J., Lee J.Y., Kweon I.S., (2018), CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference onComputer Vision (ECCV),* Munich/Germany, pp. 3-19.

[18] Xiong X., (2018), *Research on Field Rice Panicle Segmentation and Nondestructive Yield Prediction Based on Deep Learning (基于深度学习的大田水稻稻穗分割及无损产量预估研究),* PhD dissertation, Huazhong University of Science and Technology, Hubei/China.

[19] Yang W., Duan L., Yang W., (2021), Deep Learning-based Extraction of Rice Phenotypic Characteristics and Prediction of Rice Panicle Weight (基于深度学习的水稻表型特征提取和穗质量预测研究), *Journal of Huazhong Agricultural University*, 40(01), 227-235.

[20] Yang S., Wang S., Wang P., Ning Z., Xi Y., (2022), Detecting Wheat Ears Per Unit Area Using an Improved YOLOX (改进 YOLOX 检测单位面积麦穗), *Transactions of the Chinese Society of Agricultural Engineering*, 38(15), 143-149.

[21] Zhang Q., Hu S., Shu W., Cheng H., (2021), Wheat Spikes Detection Method Based on Pyramidal Network of Attention Mechanism (基于注意力机制金字塔网络的麦穗检测方法), *Transactions of the Chinese Society for Agricultura Machinery*, 52(11), 253-262.

[22] Zhang L., Chen Y., Li Y., Ma J., Du K., (2019), Detection and Counting System for Winter Wheat Ears Based on Convolutional Neural Network (基于卷积神经网络的冬小麦麦穗检测计数系统), *Transactions of the Chinese Society for Agricultural Machinery*, 50(03), 144-150.

[23] Zhang Y., Xiao D., Chen H., Liu Y., (2021), Rice Panicle Detection Method Based on Improved Faster R-CNN (基于改进 Faster R-CNN 的水稻稻穗检测方法), *Transactions of the Chinese Society for Agricultural Machinery*, 52(08), 231-240.

[24] Zhang Y., Ren W., Zhang Z., Jia Z., Wang L., Tan T., (2022), Focal and Efficient IOU Loss for Accurate Bounding Box Regression, *Neurocomputing*, 506, 146-157.

[25] Zhao F., Wang K., Yuan Y., (2014), Study on Wheat Spike Identification Based on Color Features and AdaBoost Algorithm, *Crop Journal*, 2014(01), 141-144+161.