POD PEPPER TARGET DETECTION BASED ON IMPROVED YOLOv8
/
基于改进 YOLOv8 的朝天椒目标检测研究

Jiayv SHEN, Qingzhong KONG, Yanghao LIU, Na MA*)
College of Information Science and Engineering, Shanxi Agricultural University, Taigu / China
Tel: +86-13834188480; E-mail: manasxau@163.com

DOI: https://doi.org/10.35633/inmateh-74-23

Keywords: *pod pepper, machine vision, YOLOv8, attention mechanism*

ABSTRACT

Pod pepper (Capsicum annuum var. conoides), a common variety of chili pepper, poses a challenge for traditional object detection methods due to its complex morphological features and diverse types. This study focuses on the application of machine vision technology to address the issue of pod pepper object detection. Firstly, a large number of pod pepper sample images were collected, followed by data preprocessing and annotation. Subsequently, YOLOv3, YOLOv5, YOLOv6, and YOLOv8 pod pepper object detection models were established, with YOLOv8 yielding the best detection results with a mean Average Precision (mAP) value of 81.6%. Next, different attention mechanisms were incorporated into the YOLOv8 network structure, with experimental results indicating that the Triplet Attention mechanism performed the best in pod pepper object detection, achieving an mAP value of 82.5%, a 0.9% improvement over YOLOv8. To further optimize the effectiveness of the attention mechanisms, Triplet Attention was added at different positions within the YOLOv8 network. The experiment showed that the location of adding the attention mechanism significantly impacted the pod pepper detection results. When Triplet Attention was added at the 5th layer, the best detection performance was achieved, with an mAP value of 84.1%, a 2.5% improvement over the original YOLOv8. This research provides technical support for intelligent harvesting of pod pepper.

摘要

朝天椒是一种常见的辣椒品种，由于其形态特征复杂且种类较多，传统的目标检测方法在其识别方面存在一定的挑战。本研究基于机器视觉技术，针对朝天椒目标检测问题展开研究。首先，采集了大量朝天椒样本图像，并进行了数据预处理和标注整理。其次，建立了 YOLOv3、YOLOv5、YOLOv6、YOLOv8 朝天椒目标检测模型，对比不同检测模型效果，YOLOv8 检测结果最优，检测 mAP 值为 81.6%。然后在 YOLOv8 网络结构中添加不同注意力机制，实验结果表明 Triplet Attention 机制在朝天椒目标检测中表现最好，检测结果 mAP 值为 82.5%，比 YOLOv8 提升 0.9%。为了进一步验证注意力机制的效果最大优化性，将 Triplet Attention 添加到 YOLOv8 网络不同位置，试验结果表明添加注意力机制的位置对朝天椒检测结果有显著影响。当 Triplet Attention 添加到 5 层，检测效果最好，检测 mAP 值为 84.1%，相比原始 YOLOv8 提升 2.5%。该研究可为朝天椒智能采摘提供技术支持。

INTRODUCTION

In everyday cooking, chili peppers are a popular spice among consumers and have been used as an edible vegetable, flavouring, natural colouring agent and traditional medicine (*Hernández-Pérez et al., 2020*). Pepper is a good source of provitamin A; vitamins C and E; carotenoids; and phenolic compounds such as capsaicinoids, luteolin, and quercetin (*Batiha et al., 2020*). Pod pepper, a chili variety originating from tropical regions of South America, plays a crucial role in agriculture and food processing. Efficient monitoring and harvesting of Pod peppers are therefore essential in agricultural production. By selecting and cultivating high-yielding and disease-resistant varieties, both societal and economic benefits can be enhanced.

Traditional manual monitoring methods rely on visual observation by humans, which requires prolonged attention and repetitive labour. Extended monitoring periods can lead to fatigue, making the process time-consuming, labour-intensive, and subjectively inefficient. Moreover, employing manual labour is costly and lacks periodicity, while economic efficiency of crops remains a critical production factor. Additionally, the diverse and complex morphology of pod peppers grown in fields, compounded by weather conditions and foliage occlusion, objectively diminishes the depth and dimension parameters in colour image-based target detection tasks (*Li et al., 2020*), thereby reducing the accuracy of traditional detection models. Hence, there is a need for a cost-effective, field-appropriate detection method that promotes the integration of technology and agriculture.

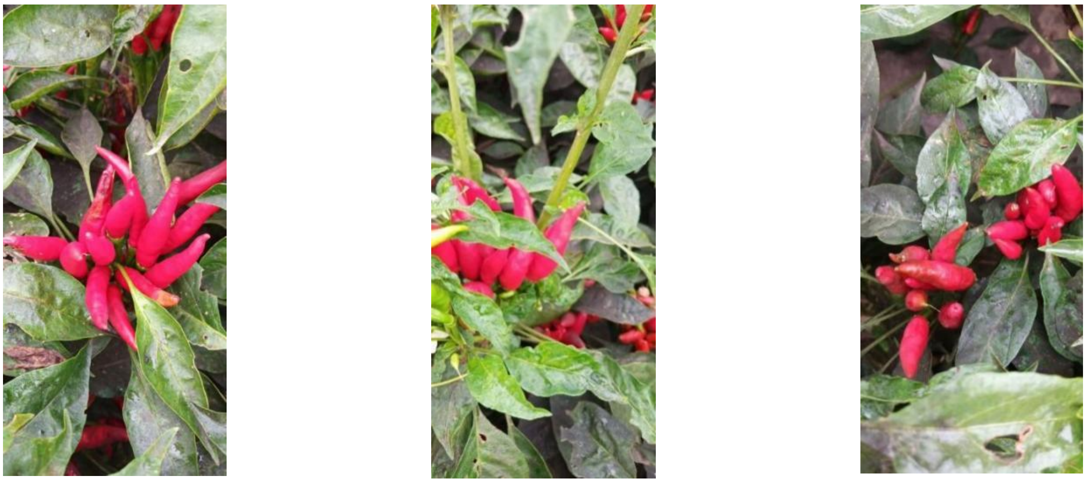
Over the past few decades, precision agriculture techniques have been richly developed to refine agricultural management practices (*Arakeri et al., 2017*). With the advancements in machine vision and deep learning technologies, automated analysis of target images using computers has rapidly progressed in fields such as crop monitoring and commercial recognition (*Cai et al., 2023*). Initially, the integration of deep learning algorithms with agriculture focused on automated sorting of agricultural products and detection and diagnosis of plant diseases and pests. Sladojevic et al. first combined the Caffe deep learning framework with the detection of plant diseases and pests, achieving an average accuracy of 96.3% in pest detection across 13 different types of leaves (*Sladojevic et al., 2016*). Kanda et al. utilized a Conditional Generative Adversarial Network, Convolutional Neural Network, and Logistic Regression to achieve an average accuracy of 96.1% in the recognition of eight plant leaf datasets. (*Kanda et al., 2021*). Farooq et al. conducted structural optimization of YOLOv8 based on SSD, YOLOv5m, Scaled YOLOv4, CenterNet, and YOLOv8m results in an average accuracy of 93.8% for recognition of small foreign object fragments (*Farooq et al., 2024*). Mahesh et al. used YOLOv3 to determine plant diseases based on symptoms on pepper leaves with a final average accuracy of 90% (*Mahesh et al., 2024*). Sapkota et al. utilized the YOLOv8 object detection and instance segmentation algorithm in conjunction with geometric shape fitting of 3D point cloud data to accurately determine the size of unripe green apples in a commercial orchard environment, with a final result of RMSE values (2.35 mm for Azure Kinect and 9.65 mm for Realsense D435i) and MAE value (1.66 mm for Azure Kinect and 7.8 mm for Realsense D435i) (*Sapkota et al., 2024*).

The aforementioned studies collectively demonstrate that object detection algorithms, particularly in the realm of agricultural modernization, show promising results in crop and fruit localization and recognition. Addressing the shortcomings of existing models in detecting pod peppers in field conditions, this study develops a pod pepper object detection model named YOLOv8-TripletAttention-5. Initially, a substantial dataset of pod pepper images was collected and annotated using labelImg. Subsequently, various YOLO series models were compared for their effectiveness in pod pepper detection, with YOLOv8 selected as the optimal base model. Different triplet attention mechanisms were then integrated at various positions on the YOLOv8 model, ultimately improving it by incorporating Triplet Attention at position 5. The results indicate a significant enhancement in efficiency and accuracy of pod pepper detection in challenging field environments where direct observation is less feasible. This research is poised to make a positive contribution to the modernization and sustainable development of agriculture.

MATERIALS AND METHODS

Data Acquisition and Pre-processing

The experimental pod pepper variety used in the study is Shengfeng Sanying No. 8. The dataset of images was collected from late July to early August 2023 at the Xiangfen County Clustered Pod Pepper Planting Base in Linfen City, Shanxi Province, China. With the purpose of this experiment is to improve the recognition accuracy of ripe pepper in the picking period, 689 images of ripe pepper were collected. The experiment captured various scenarios including obstructed and unobstructed views, as well as clear and rainy weather conditions, aiming to realistically depict the diverse situations encountered in agricultural fields during the harvest period. Figure 1 shows a selection of collected pod pepper images.



(a) shelterless (b) partial shelter (c) rainy day

Fig. 1 - Partial Data Collection

Using LabelImg for pod pepper data annotation, the YOLO format was selected. Bounding boxes were carefully drawn around each pod pepper in the images using the mouse, aiming for accuracy and completeness to ensure that no pod peppers were missed. The class label "ctj" was assigned to each annotated pod pepper. The data was then randomly split into training, validation, and test sets in a ratio of 7:2:1 and saved accordingly. Ultimately, the training set comprised 482 images, the test set contained 138 images, and the validation set consisted of 69 images.

YOLO Series Algorithm

The YOLO (You Only Look Once) series is a convolutional neural network-based object detection algorithm widely used for its fast detection speed and high accuracy. The YOLO series addresses the detection problem by processing the entire image in a single pass. Unlike traditional sliding window methods, YOLO divides the image into a grid, with each grid cell responsible for detecting objects within its boundaries and predicting multiple bounding boxes, prioritizing those whose centres fall within the cell. This eliminates the need to examine sub-regions, significantly reducing computational complexity. YOLO simultaneously predicts bounding boxes and class probabilities for objects within each grid cell, allowing for efficient and accurate detection across different images without varying window or stride sizes. The YOLO series continually introduces improvements, such as more efficient feature extraction networks and mechanisms for handling varying sizes and aspect ratios, enhancing both detection accuracy and speed.

The rise of convolutional neural networks (CNNs) has propelled the advancement of deep learning, offering greater convenience in image feature extraction. YOLO is a real-time object detection algorithm that treats object detection as a single regression problem. Therefore, it can be trained end-to-end within a single network, enabling simultaneous prediction of the positions and classes of all objects. Unlike traditional algorithms, YOLO directly divides an initial image into non-overlapping small regions and generates feature maps through convolution. Each small region of the original image corresponds to each element of the feature map, with each element predicting the objects within its respective region. YOLO features higher detection speed and a simpler network structure compared to traditional methods. The YOLO algorithm implements end-to-end object detection through an independent CNN model, utilizing multiple layers of neural networks to perform convolutions and pooling directly on images to extract essential features. This approach offers significant advantages. The mesh-like structure of the YOLO algorithm, depicted in Figure 2, includes a pair of fully connected layers and twenty-four convolutional layers. When an image enters the YOLO network, features are first extracted via convolutional networks, followed by connection through fully connected layers to produce the final predictions.

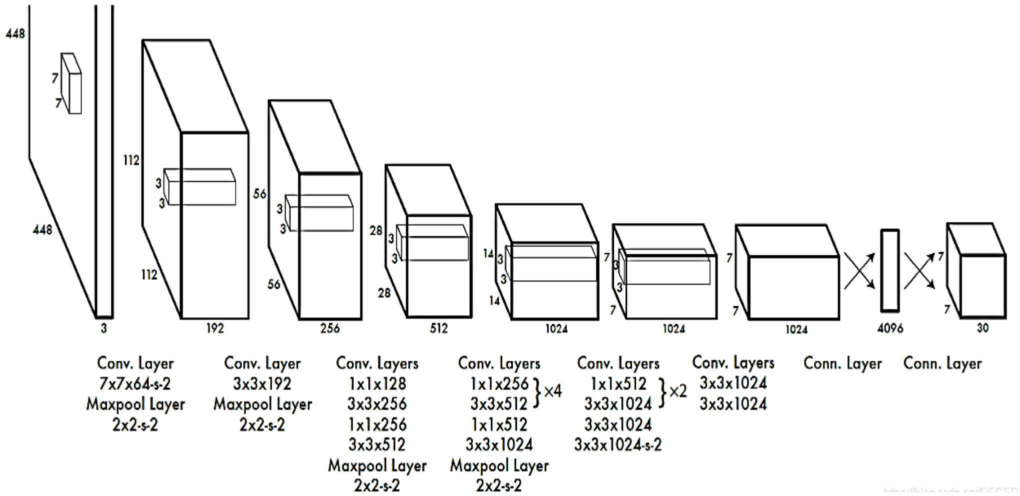


Fig. 2 - Partial network structure of the YOLO algorithm

The YOLO algorithm takes a complete image as input and outputs bounding boxes that determine the location of the target objects along with their corresponding class labels. The algorithm divides the entire image into grid cells, where each cell is responsible for predicting multiple bounding boxes and their associated class probabilities. It also predicts a confidence score for each bounding box, indicating the likelihood that the box contains an object and how confident this prediction is. The probability that an object belongs to each class within the predicted bounding box is represented by class probabilities. By setting a threshold, predictions with probabilities below the threshold can be filtered out to enhance detection accuracy.

YOLO excels in fast detection speed, ease of implementation and training, and performs well in detecting small objects. Additionally, it supports joint training with other tasks such as classification and segmentation, making it highly versatile for various applications.

YOLOv8 Model

YOLOv8 is an advanced object detection model built upon the foundation of YOLOv5, offering a new state-of-the-art (SOTA) model that further enhances performance and flexibility. The architecture of YOLOv8 is illustrated in Figure 3.

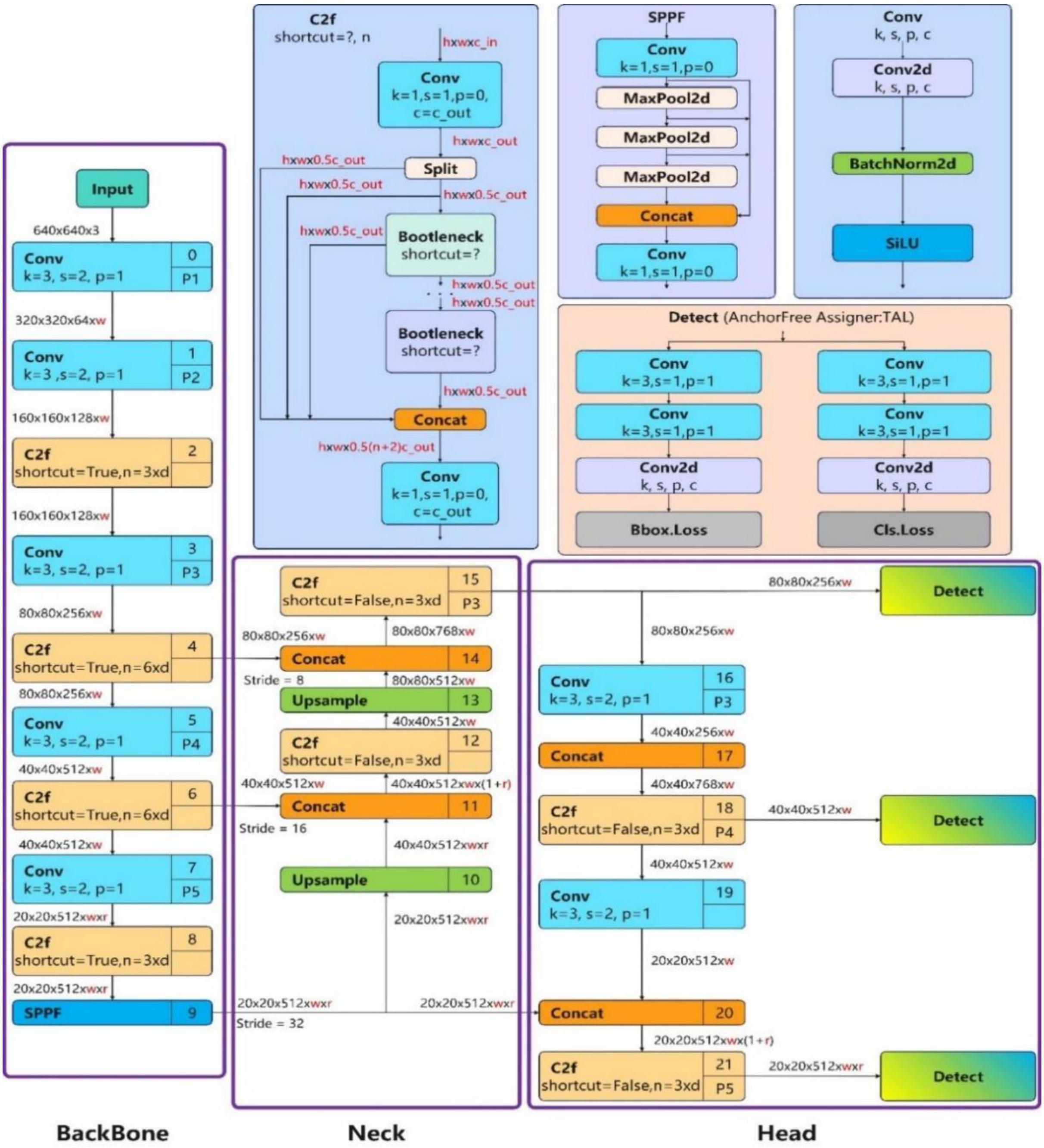


Fig. 3 –YOLOv8 Network Structure

Drawing from the design principles of YOLOv7 ELAN, YOLOv8 improves upon YOLOv5 by replacing the C3 structure with the more gradient-rich C2f structure in its backbone network and Neck section. It adjusts channel numbers for different scale models, significantly boosting model performance. YOLOv8 also introduces additional functionalities including image classification, object detection, and instance segmentation tasks. It supports command-line operation for model prediction, facilitating convenient testing and application. This study adopts YOLOv8 as the foundational algorithm for investigating target detection of Capsicum annuum var. conoides.

Triplet Attention

The Triplet Attention mechanism is specifically designed for handling sequential data. It extends traditional bidirectional attention mechanisms by enabling models to consider information from the past, present, and future simultaneously when computing attention weights. The Triplet Attention mechanism consists of three independent attention weight vectors: one each for past, current, and future information importance. These three attention weights are then combined to form the final attention weights.

The Triplet Attention mechanism offers advantages such as comprehensive context understanding, reduced risks of information leakage and overfitting, and enhanced predictive performance. Therefore, this study will primarily focus on improving the YOLOv8 algorithm based on the Triplet Attention mechanism. Figure 4 illustrates the specific implementation process of Triplet Attention, which includes three branches: upper branch, middle branch, and lower branch.

The upper branch is utilized to compute attention weights for both the channel dimension (C) and the spatial dimension (W). This branch performs a Z-Pool operation on the input tensor, followed by a convolutional layer (Conv), and finally generates attention weights using the Sigmoid function.

The middle branch is employed to capture dependencies between the channel dimension (C) and the spatial dimensions (H) and (W). This branch first undergoes identical Z-Pool and convolutional operations, followed by the generation of attention weights using the Sigmoid function in a similar manner.

The lower branch is utilized to capture dependencies among the spatial dimensions. This branch performs Z-Pool and convolution operations without altering the input, followed by generating attention weights using the Sigmoid function.

The role of the Sigmoid function in each of the three branches is to perform logistic regression (scaling the results between 0 and 1 based on a multivariate linear regression foundation). By categorizing based on a midpoint of 0.5, the essence of the classifier is to identify boundaries. Therefore, when using 0.5 as the threshold, the solution of $\hat{y} = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} = 0.5$ is sought, which is the solution at $z = \theta^T x = 0$ 。

After each branch generates attention weights, they are applied to the input, followed by aggregating the outputs of the three branches through averaging to obtain the triple attention outputs.

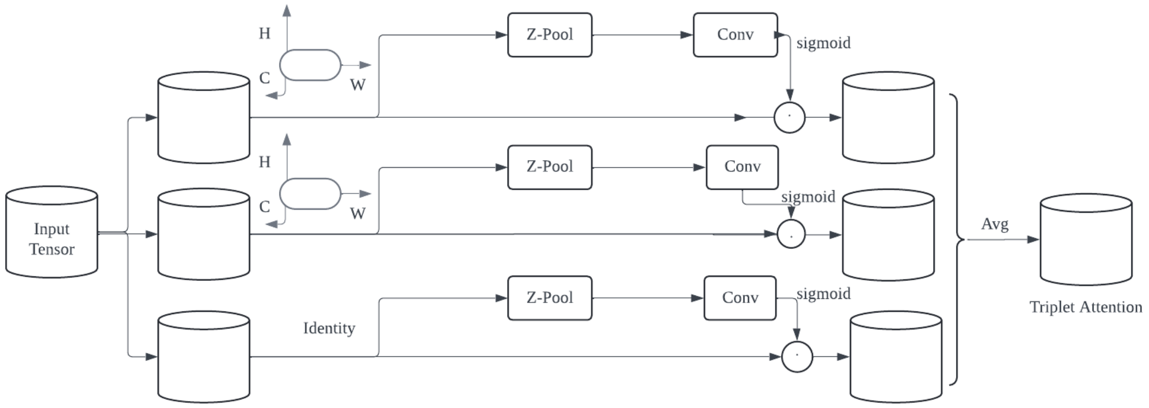


Fig. 4 –Triplet Attention Flowchart

Improved Methodology

In the context of field cultivation environments, the detection of pod pepper is influenced by stems, leaves, and weather conditions. In order to improve the accuracy of the model in the case of limited resources such as embedded devices and mobile applications, the YOLOv8n version of the YOLOv8 model is chosen, and improvements are made based on it. Attention mechanisms have been widely applied in various research fields and have shown potential to enhance traditional model detection performance (*Yadav et al., 2023*). Therefore, five different attention mechanisms were integrated —SimAM (*Yang et al., 2021*), DAttention (*Xia et al., 2022*), CPCA (*Huang et al., 2023*), SegNext_Attention (*Guo et al., 2022*), and Triplet Attention (*Misra et al., 2021*) into the 10th layer of the basic YOLOv8n model. Among these, Triplet Attention demonstrated the most promising results, prompting further testing by integrating it at different positions within the YOLOv8n model's backbone network to optimize detection performance. The experiments revealed that integrating Triplet Attention into the fifth layer of the backbone network yielded the best detection results. Figure 5 illustrates the modified architecture of the enhanced YOLOv8n model.

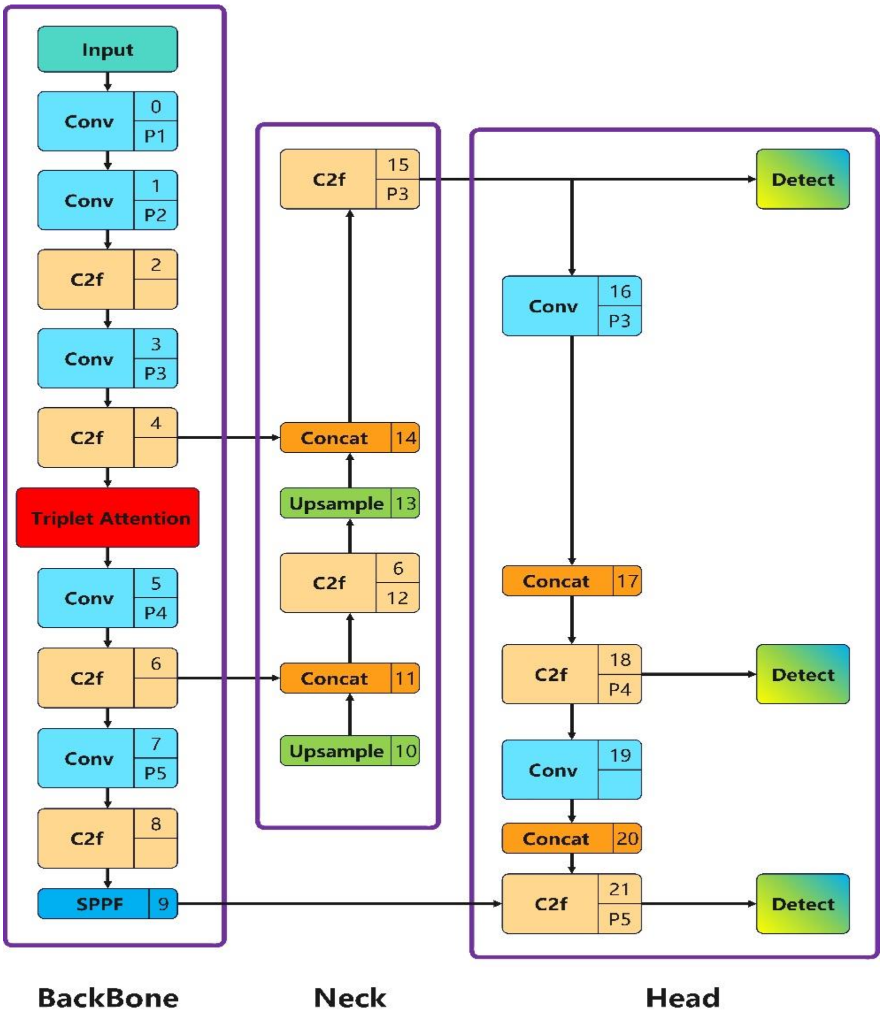


Fig. 5 –Improved YOLOv8 structure diagram

Model Evaluation Metrics

To quantitatively evaluate the performance of the proposed method against other comparative methods, this paper adopts three metrics as the evaluation standards for object detection: precision, recall, and mean average precision (mAP). These metrics can be computed using formulas 1 to 4.

precision calculation formula is:

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \tag{1}$$

recall calculation formula is:

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \tag{2}$$

mAP calculation formula is:

$$\mathbf{AP} = \sum\nolimits_{\mathbf{n}} (\mathbf{R_n} - \mathbf{R_{n-1}})\mathbf{P_n} \tag{3}$$

$$\mathbf{mAP} = \frac{\mathbf{1}}{|\mathbf{C}|}\sum\nolimits_{\mathbf{i=1}}^{|\mathbf{C}|} \mathbf{AP_i} \tag{4}$$

In this context, TP and FP represent the numbers of correct and incorrect identifications, respectively. FN represents the number of actual cases that the network model failed to detect. n is the index of data points sorted in ascending order of recall rate. $\boldsymbol{P_n}$ is the precision of the data point indexed by n, $\boldsymbol{R_n}$ is the recall of the data point indexed by n, $|\boldsymbol{C}|$ represents the number of classes, $\boldsymbol{AP_i}$ represents the mean precision of class i.

RESULTS

The operating system used for the experiment was Windows 11. The GPU model was NVIDIA GeForce RTX 3050. The CPU model was 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz. The system memory was 16GB, and the solid-state drive capacity 1TB.

The GPU acceleration libraries used were CUDA 12.3 and cuDNN 8.7. The Python version used was Python 3.11.7, and the deep learning framework PyTorch 2.3.1. The image size for deep learning training was 640×640 pixels, with 150 training epochs.

Comparison of detection results by model

As a category of rapid and efficient object detection algorithms, the YOLO series exhibits fast convergence during the detection process. Specifically, the loss functions of YOLOv3, YOLOv5, and YOLOv6 models stabilize around 100 epochs, while the YOLOv8 model stabilizes around 50 epochs. However, the Precision, Recall, and mAP values of YOLOv3, YOLOv5, and YOLOv6 models fluctuate significantly even when stabilized, resulting in relatively low accuracy, with only some achieving above 80%. In contrast, the Precision, Recall, and mAP values of the YOLOv8 model exhibit minimal fluctuations after stabilization and consistently achieve high accuracy, mostly exceeding 80%.

Table 1 summarizes the detection results for different YOLO models. Notably, YOLOv8 performed best with a precision of 77.9%, mAP of 81.6%, and mPA0.5:0.95 of 56.6%. These results indicate that YOLOv8 demonstrated superior detection performance for this experiment. Therefore, this study is based on YOLOv8 and further improves this model for the detection of Capsicum annuum var. conoides.

Table 1

Comparison of experimental parameters across models

Order	Model	Precision（%）	Recall（%）	mAP（%）	$mAP_{0.5:0.95}$（%）
1	YOLOv3	78.6	72.8	79.2	55.3
2	YOLOv5	77.9	74.1	81.5	55.5
3	YOLOv6	76.5	72.8	81.3	56.5
4	YOLOv8	77.9	71	81.6	56.6

Improved Algorithm Performance

The study attempted to enhance YOLOv8 using various attention mechanisms, added to the 10th layer of the base YOLOv8 model, with improvements shown in Table 2. From Table 2, it can be observed that after adding five different attention mechanisms, the Triplet Attention mechanism resulted in an improvement of 6.3% in Recall, a 0.9% increase in mAP, and a 1.4% increase in $mAP_{0.5:0.95}$, demonstrating a superior overall enhancement compared to the other four mechanisms.

Table 2

Different Attention Mechanisms Improve Detection Results

Order	Model	Precision（%）	Recall（%）	mAP（%）	$mAP_{0.5:0.95}$（%）
1	YOLOv8	77.9	71	81.6	56.6
2	SimAM	75.3	79.8	81.8	56.8
3	DAttention	74.1	72.8	80.4	56.5
4	CPCA	79.2	70.5	80	56.4
5	SegNext_Attention	79.7	74.1	81.6	57.5
6	Triplet Attention	74.8	77.3	82.5	58

The application of Triplet Attention can enhance the performance of the YOLOv8 algorithm, but its specific effects depend on where it is applied and the specific requirements of the task. During experiments, it is essential to carefully evaluate its effects at different positions and select the most suitable configuration based on practical needs. Therefore, to further determine the advantages of integrating the Triplet Attention mechanism into YOLOv8 for pod pepper target detection, this study conducted experiments with Triplet Attention integrated at different positions within the YOLOv8 algorithm's backbone network. The detection results are presented in Table 3, where YOLOv8-TripletAttention-x denotes the specific position where Triplet Attention is added in the YOLOv8 network.

Table 3

Triplet Attention detection results at different locations

Order	Model	Precision（%）	Recall（%）	mAP（%）	$mAP_{0.5:0.95}$（%）
1	YOLOv8-TripletAttention-2	76.7	76.6	83.7	58.3
2	YOLOv8-TripletAttention-3	75.9	79.1	82.8	57
3	YOLOv8-TripletAttention-4	85	68.8	82.8	58.5
4	YOLOv8-TripletAttention-5	79.7	76.6	84.1	58.9

Order	Model	Precision（%）	Recall（%）	mAP（%）	$mAP_{0.5:0.95}$（%）
5	YOLOv8-TripletAttention-6	78.5	74.8	82.1	56.7
6	YOLOv8-TripletAttention-7	81.6	75.1	83.9	58.4
7	YOLOv8-TripletAttention-8	76.5	77.1	82.1	56.2
8	YOLOv8-TripletAttention-9	79.1	74.4	81.7	56.9
9	YOLOv8-TripletAttention-10	74.8	77.3	82.5	58

Based on the experimental results comparison in Table 3, it is evident that the YOLOv8-TripletAttention-5 model achieved the highest overall accuracy, with an mAP of 84.1% and an $mAP_{0.5:0.95}$ of 58.9%. It performed best in detecting pod peppers.

This study compared the precision, recall, mAP0.5 and $mAP_{0.5:0.95}$ of the YOLOv8 model with that of the YOLOv8-TripletAttention-5 model during the process of gradual convergence as the number of training epochs increased, as shown in Fig. 6. It can be seen that when the overall results converged, the YOLOv8-TripletAttention-5 model outperformed the YOLOv8 model in the three metrics of precision, mAP0.5, and mAP0.5:0.95.

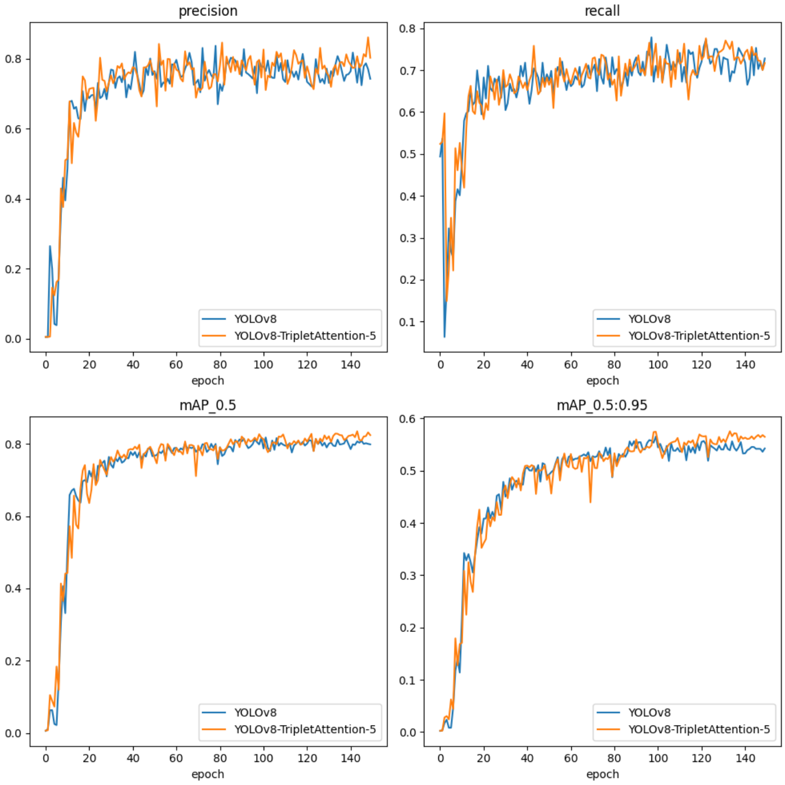


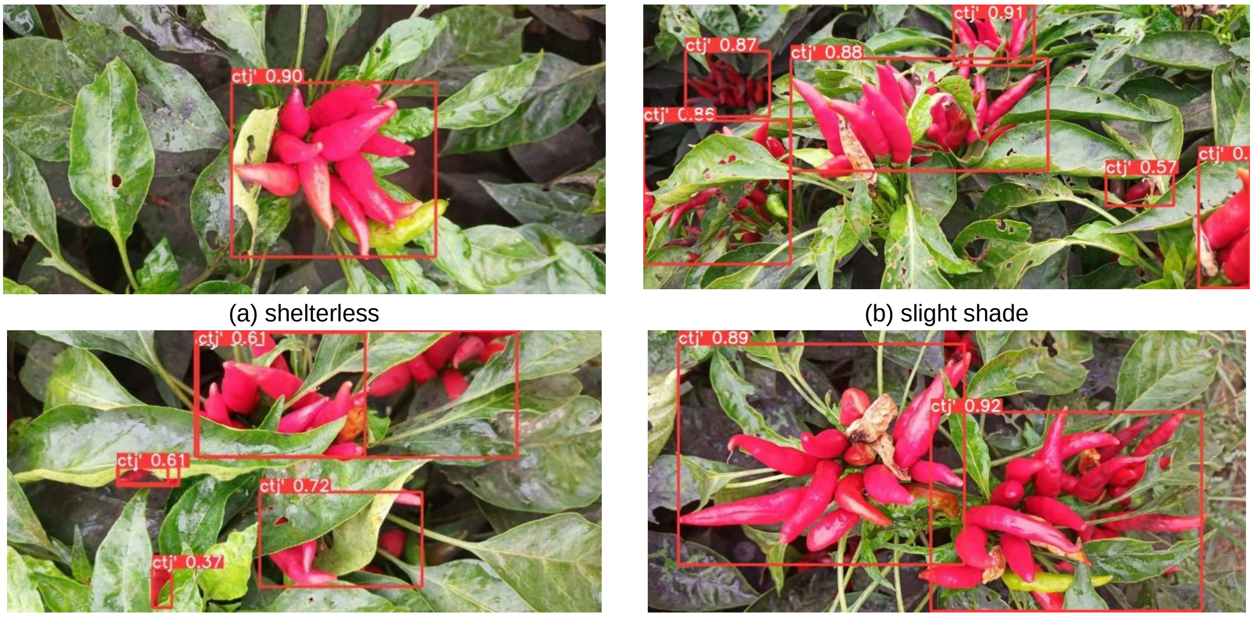
Fig. 6 – Performance curves of YOLOv8 model and YOLOv8-TripletAttention-5 model

Therefore, YOLOv8-TripletAttention-5 offers several advantages over the original YOLOv8 model: By incorporating the Triplet Attention mechanism, the model can better capture crucial information in images, thereby improving the accuracy and overall performance of object detection, specifically in detecting pod peppers. It also enhances the handling of both global and local information: the Triplet Attention mechanism balances attention between global and local information, making the model more effective in detecting targets of different scales. Moreover, it adapts better to complex backgrounds: the Triplet Attention mechanism assists the model in handling object detection tasks in complex backgrounds, thereby enhancing the model's robustness.

In summary, YOLOv8-TripletAttention-5 combines the distinct advantages of YOLOv8 and the triplet attention mechanism, complementing each other's strengths. This model shows superior performance in object detection tasks by enhancing accuracy and effectively handling complex backgrounds. Therefore, employing YOLOv8-TripletAttention-5 for the detection of Capsicum annuum var. conoides is a scientifically sound approach.

Visualisation of Test Results

The trained YOLOv8-TripletAttention-5 model was validated using pod pepper dataset, and the detection results are shown in Figure 7. From Figure7, it can be observed that our improved algorithm performs well in detecting unincluded pod pepper, partially occluded pod pepper, small-sized pod pepper targets, as well as in both sunny and rainy conditions.



(a) shelterless

(b) slight shade

(c) heavily shaded

(d) rainy day

Fig. 7 - Partial detection

CONCLUSIONS

This study successfully developed an improved object detection research based on YOLOv8, named YOLOv8-TripletAttention-5, aimed at precise localization and recognition of pod pepper positions, applicable to related production contexts.

(1) The dataset for this study was constructed under various conditions including occlusion, no occlusion, sunny, and rainy weather.

(2) Using an incremental approach to enhance the YOLO model for pod pepper detection, different YOLO models were first compared and YOLOv8 was selected for its superior performance. Subsequently, various attention mechanisms were added, with Triplet Attention proving to be the most effective. Different positions for Triplet Attention were tested to further enhance the accuracy of the improved YOLOv8 model.

(3) This study introduced Triplet Attention into the YOLOv8 model at the 5th layer of the backbone network, establishing the YOLOv8-TripletAttention-5 model. Experimental results showed an mAP (mean Average Precision) of 84.1% for pod pepper detection, a 2.5% improvement over the original YOLOv8. The mAP 0.5:0.95 achieved 58.9%, indicating enhanced precision in object detection.

(4) By integrating Triplet Attention into the original YOLOv8 model, the study potentially improved the model's ability to extract critical information from data while balancing global and partial information, thereby enhancing its effectiveness in handling complex target scenarios.

Limitations of this experiment include achieving an mAP of 84.1% for pod pepper detection, necessitating further improvements in existing object detection algorithms to enhance accuracy and robustness. Additionally, the pod pepper dataset used in this research is limited in scene diversity, highlighting the need for a more comprehensive dataset to enrich sample diversity and improve model generalization. Lastly, because this experiment is in the stage of algorithm research, when it is really applied to the actual picking activities, it may also need to further study and design lightweight models or optimization algorithms for scenes with high real-time requirements, and finally apply it to mobile devices for real-time detection and harvest of Pod pepper.

ACKNOWLEDGEMENT

This research, titled “Pod pepper target detection based on improved YOLOv8”, was funded by the Basic Research Program of Shanxi Province (202303021212115).

REFERENCES

[1] Arakeri, M. P., Kumar, B. V., Barsaiya, S., et al. (2017). Computer vision based robotic weed control system for precision agriculture[C]. *International Conference on Advances in Computing, Communications and Informatics (ICACCI).* IEEE, (pp. 1201-1205).

[2] Batiha, G. E. S., Alqahtani, A., Ojo, O. A., et al. (2020). Biological properties, bioactive constituents, and pharmacokinetics of some Capsicum spp. and capsaicinoids[J]. *International journal of molecular sciences*, *21*(15), 5179.

[3] Cai W. T., Li Z. S., Han J. N., et al. (2023). Review on application of health monitoring in pigs based on computer vision[J]. *Heilongjiang Animal Science and Veterinary Medicine*, (24):22-30.

[4] Farooq, J., Muaz, M., Khan Jadoon, K., et al. (2024). An improved YOLOv8 for foreign object debris detection with optimized architecture for small objects[J]. *Multimedia Tools and Applications*, *83*(21), 60921-60947.

[5] Guo M. H., Lu C. Z., Hou Q., et al. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation[J]. *Advances in Neural Information Processing Systems*, 35:1140-1156.

[6] Hernández‐Pérez, T., Gómez‐García, M. D. R., Valverde, M. E., et al. (2020). Capsicum annuum (hot pepper): An ancient Latin‐American crop with outstanding bioactive compounds and nutraceutical potential. A review[J]. *Comprehensive Reviews in Food Science and Food Safety*, *19*(6), 2972-2993.

[7] Huang H. J., Chen Z. G., Zou Y., et al. (2024). Channel prior convolutional attention for medical image segmentation[J]. *Computers in Biology and Medicine*, 178, 108784.

[8] Kanda, P. S., Xia, K., & Sanusi, O. H. (2021). A deep learning-based recognition technique for plant leaf classification[J]. *IEEE Access*, *9*, 162590-162613.

[9] Li Y.J., Li X.P., Zhang W.G. (2020). Survey on vision-based 3D object detection methods[J]. *Computer Engineering and Applications*, 56(01):11-24.

[10] Mahesh, T. Y., & Mathew, M. P. (2024). Detection of bacterial spot disease in bell pepper plant using YOLOv3[J]. *IETE Journal of research*, *70*(3), 2583-2590.

[11] Misra, D., Nalamada, T., Arasanipalai, A. U., et al. (2021). Rotate to attend: Convolutional triplet attention module[C]. *In Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3139–3148.

[12] Sapkota, R., Ahmed, D., Churuvija, M., et al. (2024). Immature green apple detection and sizing in commercial orchards using YOLOv8 and shape fitting techniques[J]. *IEEE Access*, *12*, 43436-43452.

[13] Sladojevic, S., Arsenovic, M., Anderla, A. et al. (2016). Deep neural networks based recognition of plant diseases by leaf image classification[J]. *Computational Intelligence and Neuroscience*, 2016(1), 3289801.

[14] Xia Z., Pan X., Song S., et al. (2022). Vision transformer with deformable attention[C]. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794-4803.

[15] Yadav, A., ＆ Vishwakarma, D. K. (2023). MRT-Net: Auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection[J]. *Expert Systems with Applications*, 232.120898.

[16] Yang L., Zhang R. Y., Li L., et al. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks[A]. *In International Conference on Machine Learning*, 139:11863-11874.