

Article

An Improved YOLOv8 Model for Lotus Seedpod Instance Segmentation in the Lotus Pond Environment

Jie Ma ^{1,*}, Yanke Zhao ², Wanpeng Fan ¹ and Jizhan Liu ³

¹ School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China; 2212107026@stmail.ujs.edu.cn

² State Key Laboratory of Intelligent Agricultural Power Equipment, Luoyang 471039, China; skliape@sina.com

³ Key Laboratory of Modern Agricultural Equipment and Technology, Jiangsu University, Zhenjiang 212013, China; 1000002048@ujs.edu.cn

* Correspondence: jma@ujs.edu.cn; Tel.: +86-0511-88791245

Abstract: Lotus seedpod maturity detection and segmentation in pond environments play a significant role in yield prediction and picking pose estimation for lotus seedpods. However, it is a great challenge to accurately detect and segment lotus seedpods due to insignificant phenotypic differences between the adjacent maturity, changing illumination, overlap, and occlusion of lotus seedpods. The existing research pays attention to lotus seedpod detection while ignoring maturity detection and segmentation problems. Therefore, a semantic segmentation dataset of lotus seedpods was created, where a copy-and-paste data augmentation tool was employed to eliminate the class-imbalanced problem and improve model generalization ability. Afterwards, an improved YOLOv8-seg model was proposed to detect and segment the maturity of lotus seedpods. In the model, the convolutional block attention module (CBAM) was embedded in the neck network to extract distinguished features of different maturity stages with negligible computation cost. Wise-Intersection over Union (WIoU) regression loss function was adopted to refine the regression inference bias and improve the bounding box prediction accuracy. The experimental results showed that the proposed YOLOv8-seg model provides an effective method for “ripe” and “overripe” lotus seedpod detection and instance segmentation, where the mean average precision of segmentation mask (mAP_{mask}) reaches 97.4% and 98.6%, respectively. In addition, the improved YOLOv8-seg exhibits high robustness and adaptability to complex illumination in a challenging environment. Comparative experiments were conducted using the proposed YOLOv8-seg and other state-of-the-art instance segmentation methods. The results showed that the improved model is superior to the Mask R-CNN and YOLACT models, with recall, precision, mAP_{box} and mAP_{mask} being 96.5%, 94.3%, 97.8%, and 98%, respectively. The average running time and weight size of the proposed model are 25.9 ms and 7.4 M, respectively. The proposed model obtained the highest mAP for lotus seedpod maturity detection and segmentation while maintaining an appropriate model size and speed. Furthermore, based on the obtained segmentation model, 3D visualization of the lotus pond scene is performed, and cloud point of lotus seedpods is generated, which provides a theoretical foundation for robot harvesting in the lotus pond.

Keywords: lotus seeds; maturity detection; segmentation; YOLOv8-seg; CBAM; WIoU



Citation: Ma, J.; Zhao, Y.; Fan, W.; Liu, J. An Improved YOLOv8 Model for Lotus Seedpod Instance Segmentation in the Lotus Pond Environment. *Agronomy* **2024**, *14*, 1325. <https://doi.org/10.3390/agronomy14061325>

Academic Editor: Gniewko Niedbala

Received: 13 May 2024

Revised: 6 June 2024

Accepted: 15 June 2024

Published: 19 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lotus seed is a popular fruit in Southeast Asia with high nutritional value. More than 80,000 mu lotus seeds have been planted in South China provinces, such as Hunan, Fujian, Jiangxi, and Hubei. The lotus seeds are grown in the lotus seedpod [1,2]. The lotus seeds in different maturity stages have different functionality, as shown in Figure 1. The ripe lotus seeds can be made into lotus seed porridge and lotus seed cuisine. The overripe lotus seeds can be used as a raw material for Chinese medicine and mooncakes. Both can bring significant economic benefits to the grower.

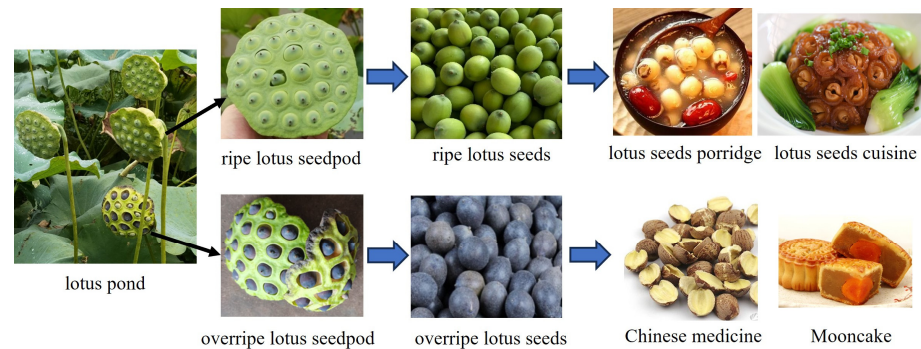


Figure 1. The functionality of ripe and overripe lotus seeds.

However, harvesting lotus seedpods is a laborious procedure, requiring substantial human resources. Between 8 and 9 h of labor are required to harvest 150–200 kg of lotus seedpods in South China. The grower needs to carry loads and walk in the lotus pond, which is a safety hazard. Additionally, China’s farm labor supply has been shrinking while farm labor demand is growing, and China’s farm labor cost is increasing. As a result, intelligent harvesting equipment is an increasingly desirable alternative for labor-intensive harvesting operations.

The accurate segmentation of lotus seedpods in different stages of maturity provides a prerequisite for developing automated harvesting equipment. However, environmental factors and growth traits create challenges in segmentation tasks, such as illumination fluctuation, occlusion of leaves and flowers, overlapping of fruits, similar background color, and the gradual change of color, texture and posture during the mature process. Fruit segmentation and detection have been studied in much literature. Traditional image segmentation techniques are generally based on image preprocessing, feature extraction, and classification. For example, Tang [3] used the color and shape characteristics of the lotus seedpods with the combination of the K-Means clustering method to segment lotus seedpods in a complex environment. Following this idea, different machine learning classifiers were used to identify the maturity of fruits based on the color and texture features of fruits [4–7]. However, images may show the variability of lighting conditions, overlap, and occlusions produced by leaves and stems that lead to the destruction of feature structure and low recognition accuracy.

To address these challenges, deep learning algorithms applying convolutional neural networks develop a series of efficient, high-accuracy, and robust models to implement classification, object detection, and instance segmentation tasks in the agricultural field [8–13]. Mask region-based convolution neural network (Mask R-CNN) combines instance segmentation into the detection network, by predicting pixel-wise masks for objects in bounding boxes [14]. Isaac et al. [15], have trained the instance segmentation model based on Mask R-CNN for unripe and ripe strawberries in images of greenhouse crops. Many improved models based on Mask R-CNN were proposed to segment different fruits (cherry tomato, apple, mango, and strawberry) [16–18]. In [13], a transformer attention mechanism was fused into the backbone of Mask R-CNN to enhance feature extraction ability. This study achieved a high mAP of 0.917 for apple segmentation. However, Mask R-CNN is composed of region proposal networks (RPN) and classification networks, which consume more computational resources and time.

The you only look once (YOLO) method regards target classification and localization as a regression problem. It has a faster detection speed without adopting RPN. Following the network architecture of YOLO, Kang and Chen [19] proposed the DaSNet-v2 model to perform detection and instance segmentation tasks for apple, where its precision value is superior to DaSNet-v1 [20]. In [12], another one-stage detector-YOLACT network [21] was used for apple pose estimation, where its inference time (94 ms) is much faster than Mask R-CNN (157 ms). The newly released model YOLOv8 was regarded as a powerful tool to generate a lightweight, high-accuracy, and fast instance segmentation model, outperform-

ing other YOLO models [22]. Since the ripe period of lotus seedpods is relatively short, a high-efficiency visual perception model is required to improve harvesting efficiency. The YOLOv8 model with improved strategies has the potential to develop an effective lotus seedpod segmentation model.

On the other hand, sufficient data is the key to training a high-accuracy model. In [8], the “copy-and-pasted” (C&P) method was adopted to generate synthetic images in order to increase the size of the dataset. Furthermore, different image processing techniques were employed to improve the performance of the segmentation model in the orchard [23] and natural scene [17]. To summarize, the combination of the C&P method and image processing methods can be an effective tool for segmentation dataset augmentation.

In this work, a visual perception framework of lotus seedpods is proposed in Figure 2, consisting of the data acquisition stage, model training and testing stage, and experiment stage. At the data acquisition stage, the collected images are annotated and augmented by means of copy-and-paste techniques and traditional data augmentation methods to address class-imbalanced problems and enhance the generalization ability of the segmentation model. A segmentation dataset is then constructed for model training and testing. Afterwards, a lotus seedpod instance segmentation model is trained based on the dataset and an improved YOLOv8-seg model, where the YOLOv8-seg model is improved by integrating the attention module and optimizing the loss function. Then, output masks of ripe and overripe lotus seedpods are obtained in 2D RGB images. At the experiment stage, a 3D point cloud image of the lotus pond scene is obtained based on RGB-D images. After point cloud image background segmentation, 2D lotus seedpod mask information is projected onto a 3D point cloud image. Three-dimensional point cloud information on ripe and overripe lotus seedpods is generated for the harvesting robot. To summarize, the main contributions of this study are as follows:

1. A class-balanced and high-quality lotus seedpod semantic segmentation dataset is constructed with a combination of copy-and-paste and traditional data augmentation methods.
2. A lightweight and high-accuracy YOLOv8-seg network incorporating a channel block attention model (CBAM) and WIoU loss function is designed to recognize ripe and overripe lotus seedpods. Ablation and comparative experiments were conducted to prove the performance of the proposed model.
3. 3D visualization of the lotus pond scene in the laboratory is performed, and cloud point information on lotus seedpods is produced through the segmentation network, which will be used to reconstruct a 3D map of a working environment and compute a proper grasping pose.

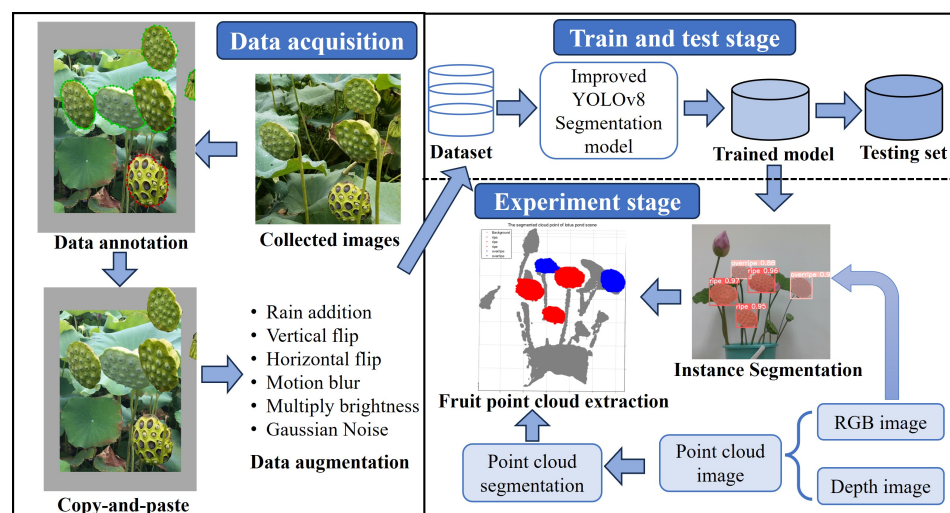


Figure 2. A perception and localization framework of lotus seedpod in different maturities.

The rest of the paper is structured as follows: Section 2 presents data acquisition and the methodology, including a description of the improved YOLOv8-seg network used for lotus seedpod instance segmentation. Section 3 evaluates the performance of the segmentation, while Section 4 discusses the results. Section 5 includes conclusions and future work.

2. Materials and Methods

2.1. Data Collection

Lotus seedpod images were collected from two lotus ponds in Huashi County, Hunan Province, China. A camera, mobile phone, and drone were used to take images of ripe and overripe lotus seedpods from different angles, with a distance of 0.6–1 m. Images were collected from the morning until the afternoon in different weather conditions, such as full sun, partly cloudy, rainy, and overcast. Figure 3a,b gives examples of acquired images. A total of 887 original images were selected to compose a representative dataset for lotus seedpod instance segmentation tasks.

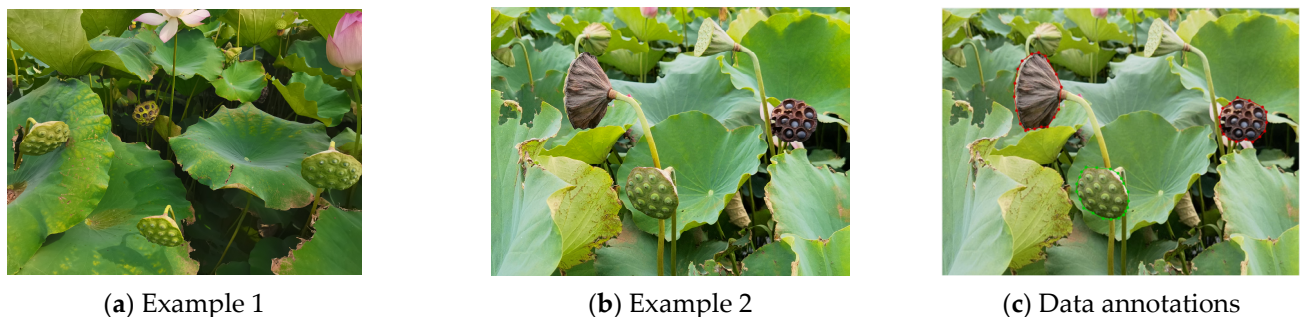


Figure 3. Data examples and annotations.

Regarding the acquired images, the contours of the ripe and overripe lotus seedpods were annotated using LabelMe 5.4.1, as shown in Figure 3c. Two classes of objects were annotated: Ripe and overripe lotus seedpods. As shown in Figure 3, lotus seeds and lotus seedpods in the ripe stage are green, and the posture of lotus seedpods is drooping. However, at the overripe stage, lotus seeds are black and lotus seedpods are brown. The posture of lotus seedpods is upright or drooping. Note that lotus seedpods with irregular shapes were not annotated.

Since the ripe stage of lotus seed is relatively short (4–5 days) compared to the overripe stage (over 7 days), the label size of the two classes is unbalanced. Copy-and-paste is an effective data augmentation strategy to increase the size of the ripe class [24]. Specifically, the ripe instances are randomly selected from one image and pasted into another image, while applying random scale jittering and random horizontal flipping. Figure 4 gives examples of the copy-and-paste strategy.

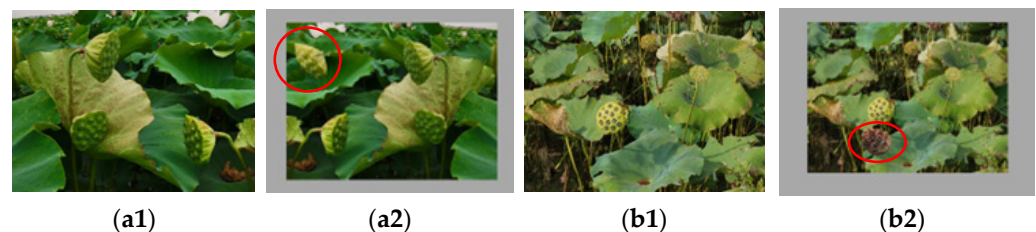


Figure 4. Examples of copy-and-paste strategy. (a1,b1) Original images. (a2) Result of copy-and-paste technique applying random scale jittering and horizontal flipping. (b2) Result of copy-and-paste technique applying random scale jittering. The red circle indicates pasted masks.

Traditional data augmentation methods are considered to enrich dataset diversity, improve the model's generalization capability, and simulate the real condition of the lotus

pond, such as weather factors, noise caused by hardware aging, and vibration generated during equipment movement [2]. These challenges will be addressed by applying the different image processing techniques shown in Figure 5. Motion blur, rainy effect, flipping, multilevel brightness and Gaussian noise are considered to mitigate the effects of the aforementioned factors. By employing traditional augmentation methods and copy-and-paste augmentation methods, the original 887 images are expanded to 2350 to create the dataset, as shown in Table 1. The established dataset is randomly divided into the training set, validation set, and testing set with a ratio of 70%, 20%, and 10%, respectively.

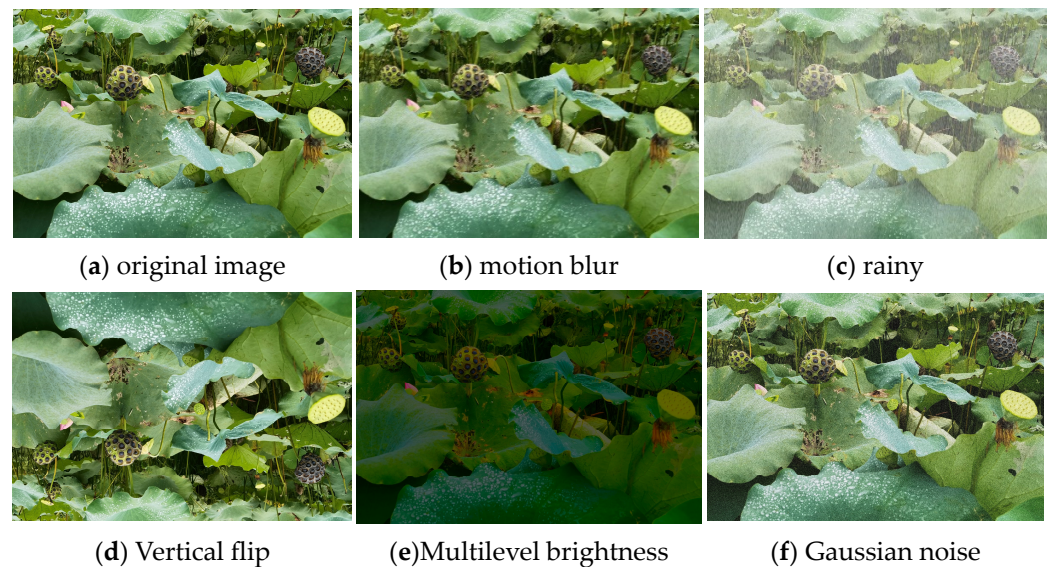


Figure 5. The examples illustrate traditional data augmentation methods.

Table 1. The composition of the lotus seedpods dataset.

	Multilevel Brightness	Vertical Flip	Horizontal Flop	Rainy	Gaussian Noise	Motion Blur	C&P	Original	Total
Num	227	110	120	240	240	230	296	887	2350

2.2. Proposed Methodology

2.2.1. YOLOv8-Seg Network

YOLOv8 follows the network architecture of YOLOv5. YOLOv8-seg is a segmentation branch of the YOLOv8 model, as shown in Figure 6 [22]. Its main components include an input module, backbone network, neck network, and prediction module. The input module implements mosaic data augmentation and adaptive image scaling. The backbone network is composed of Conv, C2f, and SPFF modules to generate feature maps of 1/8, 1/16, and 1/32 sizes of the input images. These feature maps are the output of C3, C4, and C5. The C2f block has the advantage of strong feature extraction capability. In Figure 6 (blue box), the C2f block utilizes across-layer connections to enhance gradient flow information, and split operations are introduced to increase gradient information and reduce the computation cost. The neck network adopts a PANet structure to fuse the multiscale features extracted from the backbone network.

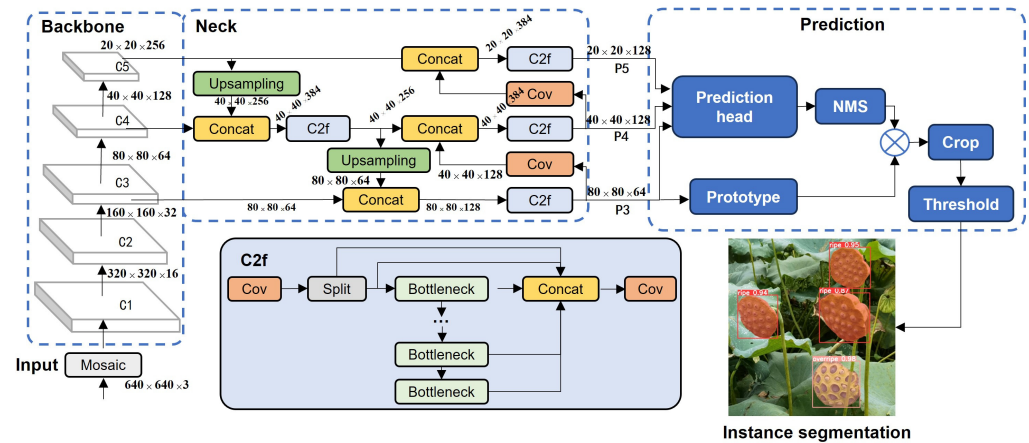


Figure 6. YOLOv8-seg network.

In the prediction part, following the idea of the YOLACT network [21], the prediction head block consists of the detection branch and the segmentation branch, which perform detection and segmentation tasks, respectively. The detection branch generates classes, bounding boxes, and mask coefficients. The segmentation branch generates prototype masks. After the non-maximum suppression (NMS) block, the predicted mask coefficients and prototype masks of each instance are assembled by matrix multiplication. The crop and threshold operations are used to construct the final instance.

2.2.2. Model Improvement

In order to improve lotus seedpod detection and segmentation accuracy and computation speed, an improved YOLOv8-seg model integrates three CBAM modules into the neck network to effectively boost the representation power of the deep learning network, as shown in Figure 7.

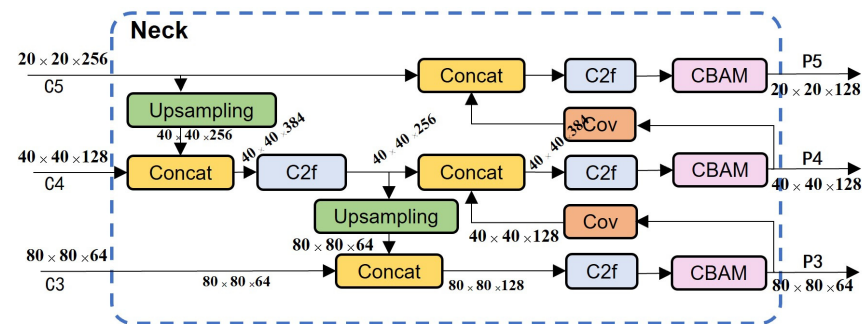


Figure 7. YOLOv8-seg neck network with improved technique.

As shown in Figure 8, CBAM consists of a channel attention module and a spatial attention module. The spatial dimension of the input feature map (F) is aggregated by using average-pooling and max-pooling operations simultaneously, where the average-pooling feature F_{avg}^c and max-pooling feature F_{max}^c are generated. Both features are forwarded to a multi-layer perceptron (MLP) network to obtain channel attention map M_c . Afterward, the output feature vectors are merged through element-wise summation. Overall, the channel attention is calculated as:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

where σ denotes the sigmoid function.

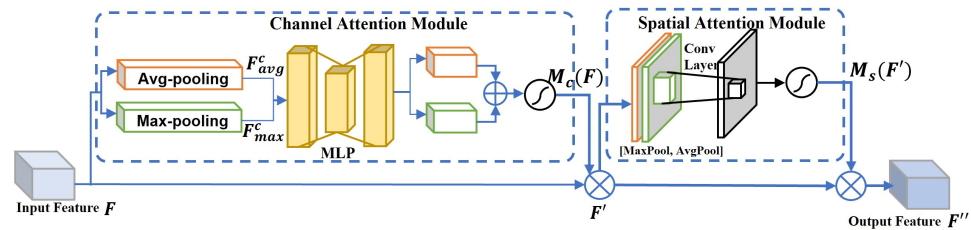


Figure 8. Schematic diagram of CBAM module.

The channel information of the input feature map (F') is aggregated along the channel axis by using average pooling and max-pooling operations, where average-pooling feature F_{avg}^s and max-pooling feature F_{max}^s are generated. F_{avg}^s and F_{max}^s are concatenated and convolved by a convolution layer to generate a spatial attention map M_s . Mathematically, the spatial attention is calculated as:

$$M_s(F') = \sigma\left(f^{7 \times 7}([AvgPool(F'); MaxPool(F')])\right) = \sigma\left(f^{7 \times 7}\left([F_{avg}^s; F_{max}^s]\right)\right) \quad (2)$$

where $f^{7 \times 7}$ denotes a convolution operation with the filter size of 7×7 . Overall, given an input feature map F , the attention process can be summarized as:

$$F' = M_c(F) \otimes F \quad (3)$$

$$F'' = M_s(F') \otimes F' \quad (4)$$

where $\otimes$ denotes element-wise multiplication. F'' is the final output feature. During the multiplication process, the channel attention information is broadcasted along the spatial dimension, while the spatial attention values are copied along the channel axis.

In the original YOLOv8-seg network, the CIoU loss function is used to locate the bounding box. However, if the ratio of height and width of the target and prediction boxes varies linearly, then the penalty term will not work and the anchor box will not be optimized. The WIoU loss function has a dynamic non-monotonic focusing mechanism to develop a wise gradient gain allocation strategy [25]. Due to the existence of low-quality images in the training data, geometric factors such as distance and aspect ratio will aggravate the penalty for low-quality examples and degrade the generalization performance of the model. Given that the anchor box coincides well with the target box, the penalty of geometric factors should be weakened by the loss function, and less intervention in training will make the model have better generalization ability. Following this idea, WIoU v1 with two attention mechanisms is defined as follows:

$$\mathcal{L}_{WIoUv1} = \mathcal{R}_{WIoU} \mathcal{L}_{IoU} \quad (5)$$

The first attention mechanism is $\mathcal{R}_{WIoU}$, which is the penalty term to amplify $\mathcal{L}_{IoU}$ of the ordinary-quality anchor box. The second attention mechanism is $\mathcal{L}_{IoU}$, which reduces $\mathcal{R}_{WIoU}$ of the high-quality anchor box and focus on the distance between central points when the anchor box coincides well with the target box. The definitions of $\mathcal{L}_{IoU}$ and $\mathcal{R}_{WIoU}$ are given as follows:

$$\mathcal{L}_{IoU} = 1 - IoU = \frac{|B \cap B_{gt}|}{|B \cup B_{gt}|} \quad (6)$$

$$\mathcal{R}_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (7)$$

where the properties of anchor box and target box are described as $B = [x, y, w, h]$, $B_{gt} = [x_{gt}, y_{gt}, w_{gt}, h_{gt}]$, respectively. W_g and H_g are the size of the smallest enclosing box, as shown in Figure 9.

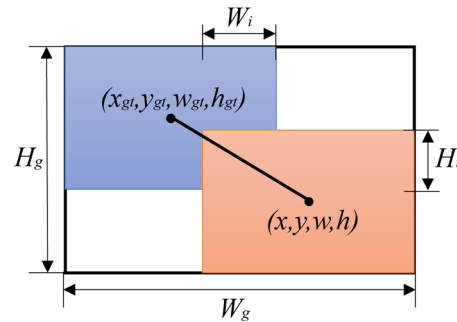


Figure 9. Geometric representation of Intersection over Union.

In order to increase the contribution of hard examples to the loss value and improve classification performance, WIoU v3 with a dynamic non-monotonic focusing mechanism is proposed based on WIoU v1. The outlier degree of anchor box β is characterized by the ratio of $\mathcal{L}_{IoU}$ to $(\overline{\mathcal{L}_{IoU}})$ to evaluate the quality of the anchor box:

$$\beta = \frac{\mathcal{L}_{IoU}}{\overline{\mathcal{L}_{IoU}}} \quad (8)$$

where $\overline{\mathcal{L}_{IoU}}$ represents the exponential running average with momentum. A lower β presents the higher quality of the anchor box, where a small gradient gain is assigned to focus on ordinary-quality anchor boxes. Furthermore, a small outlier degree will effectively prevent large harmful gradients from low-quality examples. A non-monotonic focusing coefficient using β is applied to construct WIoU v3:

$$\mathcal{L}_{WIoUv3} = r\mathcal{L}_{WIoUv1} \quad r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \quad (9)$$

where α and δ are hyperparameters. Due to $\overline{\mathcal{L}_{IoU}}$ being dynamic, the quality demarcation standard of anchor boxes is dynamic, which enables WIoU v3 to make the gradient gain allocation strategy that is most appropriate to the current situation at any moment.

2.2.3. Evaluation Metrics

To evaluate the accuracy of the lotus seedpod instance segmentation model, precision (P), recall (R), mean average precision (mAP) are utilized as performance metrics in this work [26,27]. In the case of detection and segmentation, the sets can be used to evaluate bounding boxes and masks. P measures the fraction of true positive objects in prediction, while R measures the fraction of true positive objects that are successfully detected. mAP represents the mean average precision of bounding box or pixels segmentation of all categories, which can be denoted as mAP_{box} and mAP_{mask} , respectively. The expression of the P , R , AP , and mAP are listed as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \quad (10)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (11)$$

$$AP = \int P(R) dR \times 100\% \quad (12)$$

$$mAP = \frac{1}{C} \sum_{i=1}^C AP(i) \times 100\% \quad (13)$$

where TP (true positive) denotes the number of correctly predicted lotus seedpod categories; FP (false positive) denotes the number that was incorrectly predicted as lotus seedpod categories or background; FN (false negative) represents the number of lotus seedpod

category that were not detected or misdetected as negative samples. AP measures the average precision of bounding box or pixel segmentation. C is the number of categories. In addition, this work also considers the model size and detection time to comprehensively evaluate its performances.

2.2.4. Model Parameters and Software

The training and testing for the network were conducted on a workstation with an Intel Core i7-11700K @3.6GHZ, NVIDIA GeForce RTX 3070 8G GPU, Windows 10 64-bit system. Software tools, including CUDA 11.1, cuDNN 8.2.0, and Python 3.8, were utilized. All experiments are conducted in the same environment.

The training model was initialized with pre-trained weights based on the COCO dataset. An initial learning rate of 0.01 was used. The momentum and weight decay were 0.937 and 0.0005, respectively. The training batch size was 24, and the maximum number of epochs was 150. The resolution of the input images was set to 640×640 pixels. Stochastic gradient descent (SGD) was selected as an optimizer.

3. Results

3.1. Performance of Data Augmentation Technique

To clarify the impact of data augmentation techniques on model detection and segmentation performance, the YOLOv8n-seg model was employed to evaluate experimental results based on different datasets. As shown in Table 2, the dataset optimized by the copy-and-paste (C&P) technique and traditional augmentation (TA) methods outperforms the original dataset and the dataset only optimized by C&P. Compared with the original dataset, mAP_{box} and mAP_{mask} for dataset optimized by C&P and TA both increased by 7.7%. Note that C&P effectively addresses the class-imbalanced problem, where the AP of the bounding box for the ripe class increases from 83% to 91.2%. However, the AP of the overripe class decreases by 2.6% after using C&P. It can be inferred that the expansion of the dataset makes the label class more balanced, which presents challenges for classification tasks and deteriorates the model performance. However, this problem can be solved by using the traditional augmentation method to generate a bigger dataset, where the box AP of the overripe class increased by 5.6%.

Table 2. Comparison of performance on instance segmentation of lotus seedpods among different data augmentation technique.

	Type	Box			Mask			mAP_{box} (%)	mAP_{mask} (%)
		P (%)	R (%)	AP (%)	P (%)	R (%)	AP (%)		
None	ripe	78.5	81.8	83	78.5	81.8	83	89.7	89.7
	overripe	88.7	95.2	95.2	88.7	95.2	96.4		
+C&P	ripe	87.2	85.5	91.2	87.2	85.5	91	91.9	91.6
	overripe	92.3	85.6	92.6	92.3	85.6	92.3		
+C&P + TA	ripe	94.9	92	96.6	94.9	92	96.6	97.4	97.4
	overripe	97.3	95.1	98.2	97.3	95.1	98.1		

3.2. Ablation Experiment and Robustness Test

To clarify the impact of network improvement on the detection and segmentation performance of the model, this paper applied the YOLOv8n-seg network as the baseline model, and the improved components were added in turn through ablation experiments to verify the necessity of the corresponding improvement.

As shown in Table 3, adding the CBAM module to the neck network helps to extract abundant semantic features in the deep network, resulting in a 0.3% increase in mAP_{box} value and a 0.2% increase in mAP_{mask} value. Based on this, the WIoU loss function was

introduced to improve the mAP_{box} and mAP_{mask} values by 0.3%, respectively. Additionally, P and R were increased by 0.4% and 0.6%, respectively. Overall, the improved YOLOv8-seg model (YOLOv8n-seg + CBAM + WIoU) showed significant improved detection and segmentation performance compared to the baseline model.

Table 3. Comparison results of ablation experiment performance.

Models	P (%)	R (%)	mAP_{box} (%)	mAP_{mask} (%)	Time	Size
YOLOv8n-seg	96.1	93.5	97.4	97.4	26.9 ms	6.8 MB
+CBAM	96.1	93.7	97.7	97.6	25 ms	7.4 MB
+CBAM + WIoU	96.5	94.3	98	97.9	25.9 ms	7.4 MB

In order to verify the robustness of the proposed model, the instance segmentation performance of lotus seedpods is tested in a complex nature environment, as described in Figure 10. As shown in Figure 10a, the proposed model can segment overlapping targets with high accuracy. Figure 10b–d shows that the proposed model has strong robustness to segment ripe and overripe lotus seedpods under different illumination conditions (backlight, overexposure, and shadow light). In Figure 10d, far and small lotus seedpods can be classified correctly and segmented accurately by the proposed model without missed or false detection. Moreover, the harvesting efficiency can be greatly improved by recognizing far or small targets in the field of view.

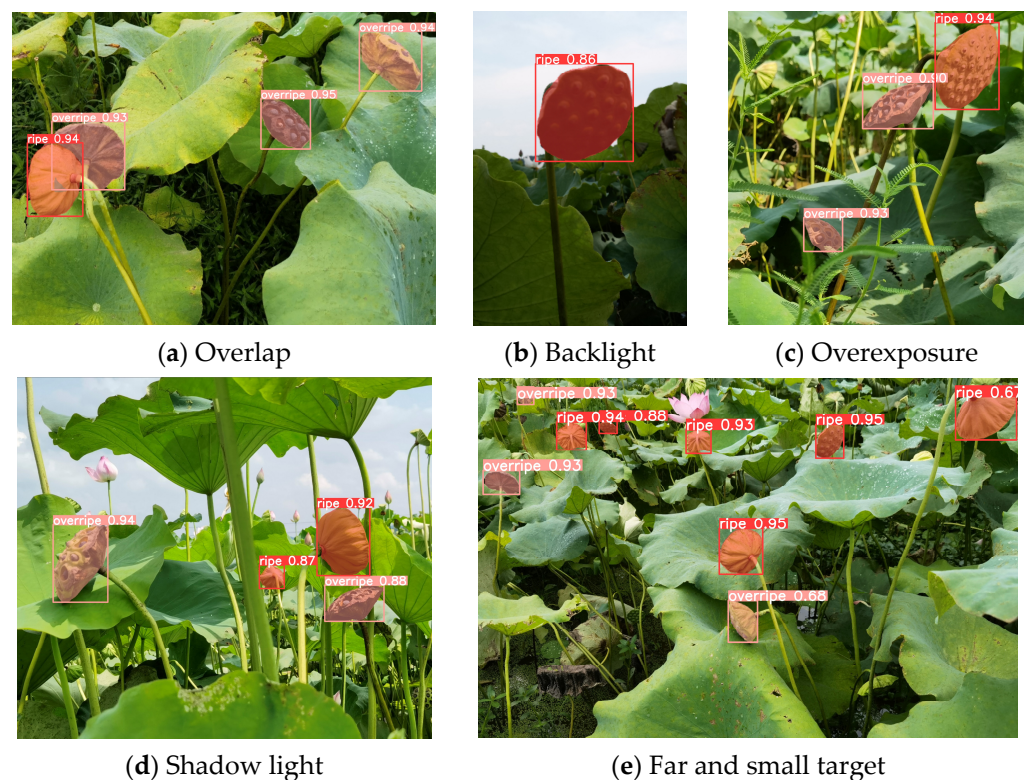


Figure 10. Lotus seedpods segmentation results under different conditions.

3.3. Comparison to State-of-the-Art Methods

A series of experiments were conducted to compare the performance of improved YOLOv8-seg, YOLOv8n-seg, YOLOv5, Mask R-CNN, and YOLACT. YOLACT and Mask R-CNN are both widely used segmentation detectors. The performance metrics are shown in Table 4 and Figure 11. From Table 4, it is shown that the proposed model is superior to YOLACT, Mask R-CNN with backbone ResNet 50, and Mask R-CNN with backbone ResNet 101 in accuracy, speed and size, with mAP_{box} being 98%, mAP_{mask} being 97.9%,

detection time being 25.9 ms, and model size being 7.4 MB. In terms of mAP value, the proposed model has a higher mAP value and faster detection speed compared with the YOLOv8n-seg and YOLOv5 models in [2], with an appropriate model size.

Table 4. Comparison of performance on instance segmentation of lotus seedpods among different models.

Models	mAP_{box}	mAP_{mask}	Time	Size
Proposed method	98%	97.9%	25.9 ms	7.4 MB
YOLOv8n-seg	97.4%	97.4%	26.9 ms	6.8 MB
YOLOv5 [2]	95.8%	95.7%	97.9 ms	14.8 MB
YOLACT	93.4%	93.3%	204 ms	189 MB
Mask R-CNN (ResNet 50)	95.7%	95.7%	289 ms	348 MB
Mask R-CNN (ResNet 101)	95.4%	95.4%	296 ms	496 MB

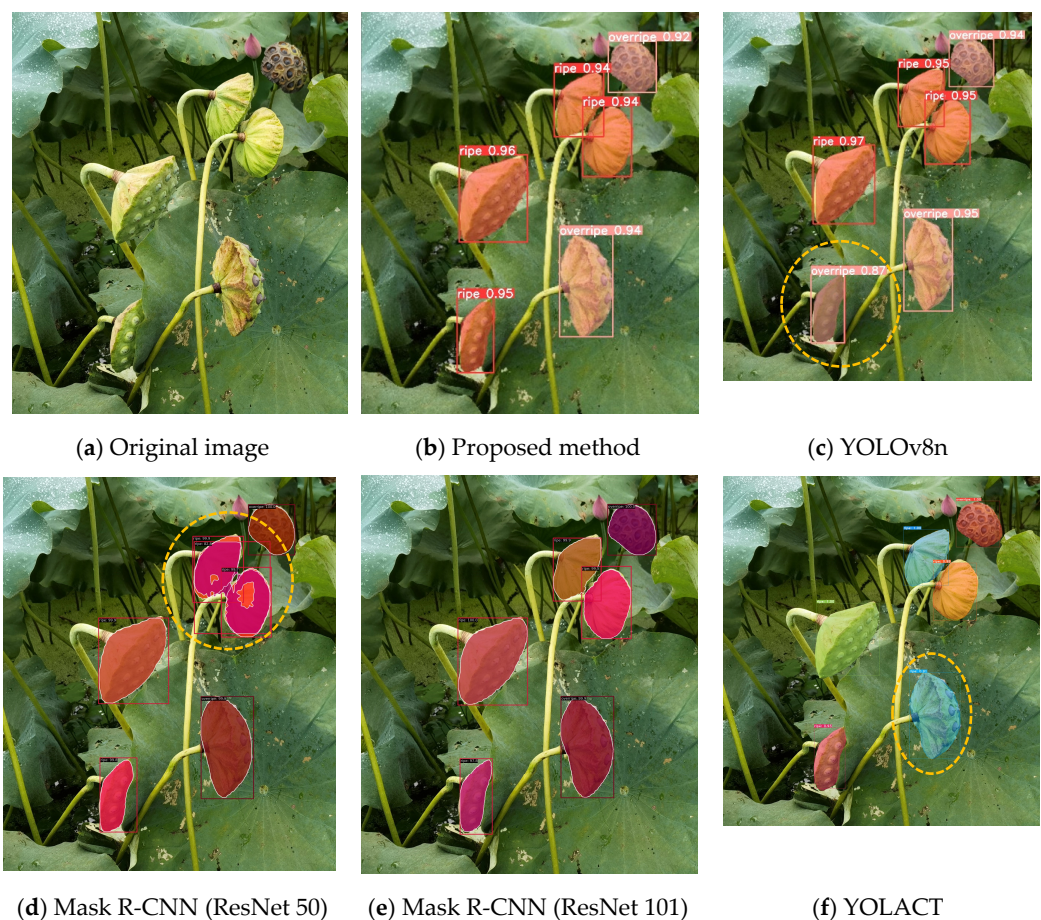


Figure 11. Lotus seedpods segmentation results based on different deep learning networks.

Figure 11a shows an image of a lotus pond scene in which lotus seedpods have different view angles (front, side, and back view), maturities (ripe and overripe), and different degrees of occlusion. Our model achieved accurate segmentation of ripe and overripe lotus seedpods, back or side-view lotus seedpods, and occluded lotus seedpods without false or missed detections, as shown in Figure 11b, whereas the Mask R-CNN with backbone ResNet 101 model has the same performance, as shown in Figure 11e. However, the model size of Mask R-CNN (ResNet 101) is much bigger than the proposed model, as described in Table 4. The YOLOv8n-seg and YOLACT models both have a false segmentation target, as indicated in Figure 11c,f, with a yellow dotted circle. Mask R-CNN (ResNet 50) has a redundant detection target, as shown in Figure 11d with a yellow dotted circle.

3.4. 3D Visualization of Lotus Pond Scene in Laboratory

A vision-perception platform of lotus seedpods was established in the laboratory, as shown in Figure 12. On the platform, a depth camera (Intel Realsense D435) collected an RGB-D image of the artificial lotus pond scene and sent it to the mobile controller (Raspberry Pi 4B). The RGB and depth images are displayed in the monitor. The collected RGB-D images are processed using the proposed segmentation model and visualized in MATLAB. As shown in Figure 13, a 3D point cloud is generated by RGB-D images. The 2D-segmented mask information in the RGB image is projected into the 3D point cloud image. In the end, a 3D point cloud image with semantic information can clearly describe the working environment of the harvesting robot in a lotus pond. The obtained information is the basis for constructing a 3D map of working spaces and computing the pose of each lotus seedpod in order to improve the success rate and reduce the rate of fruit damage in the harvesting process. These works will be involved in future work that develops an intelligent robotic system for lotus seedpod harvesting.

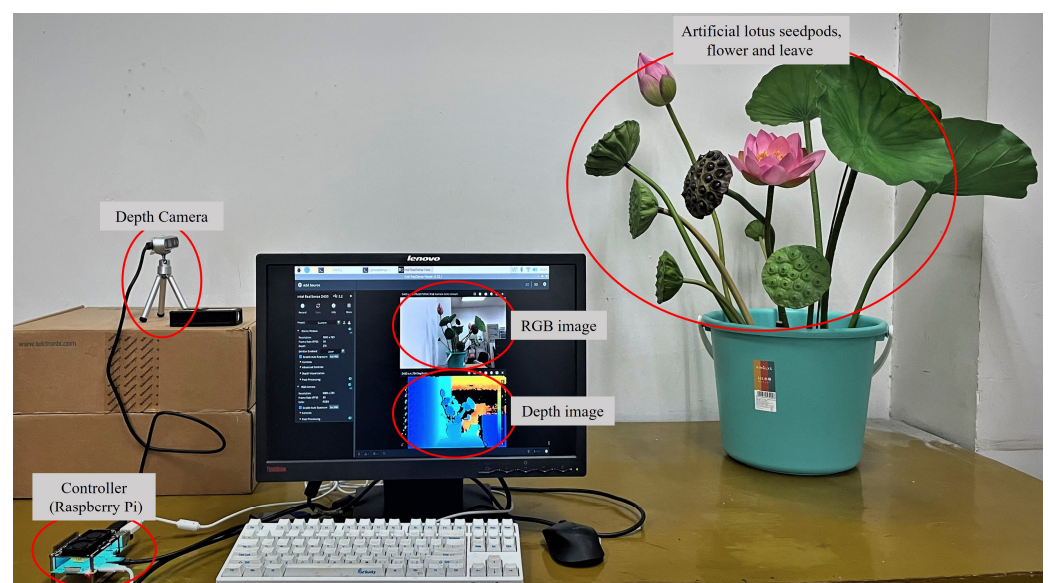


Figure 12. Lotus seedpods vision perception experiment platform.

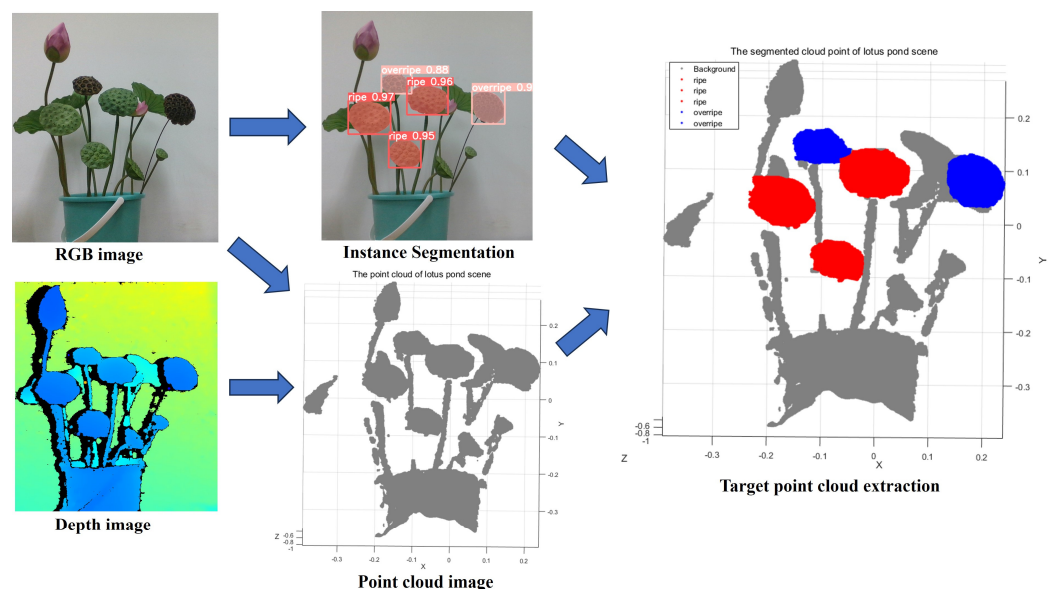


Figure 13. Point cloud extraction and segmentation procedure of lotus pond scene.

4. Discussion

In the agricultural field, many studies have been carried out on the design of algorithms for the segmentation of different fruits in images. As for lotus seedpods, our previous work dealt with the detection of overripe lotus seedpods in images in order to implement this in an automatic harvesting machine [2]. However, because lotus seedpods are an upwardly growing conical fruit with different poses, the harvesting system needs segmentation with a high level of detail to determine the grasping pose of the end effector.

In practice, the growth stage of lotus seedpods in a lotus pond is inconsistent. The ripe stage of lotus seedpods is only 5–7 days, which makes it difficult to collect sufficient data. On the other hand, the lotus seedpods at the unripe and ripe stages have similar features, such as: The color of lotus seeds is green. The ripe lotus seedpods are easy to be misdetected. Therefore, accurately detecting and segmenting ripe lotus seedpods is a great challenge. We addressed this problem by using the data generation method. By segmenting and copying the ripe lotus seedpod from the original image and pasting it to randomly selected images, we were able to generate a large number of synthetic images from the original images and overcome the class-imbalanced problem. The synthetic images and original images were further augmented using horizontal flip, vertical flip, multiply brightness, gaussian noise, rainy effect, and motion blur methods to create a large training dataset for deep learning. Table 2 shows that the C&P method effectively improves the average precision of ripe lotus seedpods, while traditional data augmentation methods increase the average precision of both classes.

In this paper, we presented an improved YOLOv8-seg model to perform instance segmentation of lotus seedpods. An improved YOLOv8-seg model has been created by introducing a lightweight CBAM module and WIoU loss function. The CBAM module adaptively refines features in the YOLOv8-seg network, which allows the model to focus on the critical context information of lotus seedpods and enhances its capacity to extract distinguished features with negligible computation overheads. The WIoU loss function provides a gradient gain allocation strategy to dynamically balance the effects of high-quality and low-quality anchor boxes and improve the detector's overall performance. The improved model has higher mAP values, faster computational speed, a lighter model size, and higher segmentation quality than the Mask R-CNN and YOLACT models, as demonstrated in Figure 11 and Table 4. The prediction for ripe and overripe classes was over 95% accurate, indicating that instance segmentation networks can be a reliable tool for making autonomous harvesting decisions. In addition, the proposed model has demonstrated its segmentation performance in a 3D point cloud image based on a visual perception platform for a lotus pond scene in the laboratory. The proposed model successfully detects and segments “ripe” and “overripe” lotus seedpods, whereas only overripe lotus seedpods were detected in reference [2]. Furthermore, references [28,29] realized the detection and segmentation of lotus seedpods with high mAP values, respectively. However, they did not distinguish the maturity stage of lotus seedpods.

5. Conclusions

In this study, the YOLOv8-seg model was improved by incorporating the CBAM module and WIoU loss function in order to implement instance segmentation tasks for ripe and overripe lotus seedpods in an unstructured environment. Experiment results show that the proposed YOLOv8-seg model is superior to state-of-the-art models, such as Mask R-CNN and YOLACT, in speed, accuracy, and model size. The recall, precision, mAP_{box} , and mAP_{mask} of the proposed model were 96.5%, 94.3%, 98%, and 97.9%, respectively. The weight size and the average running time to process an image were 7.4 M and 25.9 ms, respectively. The proposed model has good robustness and effectiveness for the visual perception of the lotus pond. In addition, a visual sensing platform of a lotus pond was established in the laboratory, where a 3D cloud point map with semantic information was obtained to describe the working environment of the harvesting robot.

Future work will focus on applying the segmentation model to develop a lotus seedpod pose reconstruction algorithm. The corresponding control strategy for harvesting robot end-effectors will also be considered.

Author Contributions: Conceptualization, J.M. and W.F. methodology, W.F.; software, J.M.; validation, J.M., Y.Z., W.F. and J.L.; formal analysis, J.L.; investigation, J.M.; resources, W.F.; data curation, J.M.; writing—original draft preparation, J.M. and J.L.; writing—review and editing, J.M. and W.F.; visualization, J.L.; supervision, Y.Z.; funding acquisition, J.M., Y.Z., W.F. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by “China Postdoctoral Science Foundation”, China (grand No. 2023M731374), “Open Fund of State Key Laboratory of Intelligent Agricultural Power Equipment”, China (grant No. SKLIAPE2023010), “Jiangsu University Senior Talents Start-up Fund”, China (grant No. 5501140007), “Jiangsu Province Innovation and Entrepreneurship Doctoral Program”, China (grant No. JSSCBS20221205).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Chen, C.; Li, G.; Zhu, F. A novel starch from lotus (*Nelumbo nucifera*) seeds: Composition, structure, properties and modifications. *Food Hydrocoll.* **2021**, *120*, 106899. [\[CrossRef\]](#)
- Ma, J.; Lu, A.; Chen, C.; Ma, X.; Ma, Q. YOLOv5-lotus an efficient object detection method for lotus seedpod in a natural environment. *Comput. Electron. Agric.* **2023**, *203*, 107635. [\[CrossRef\]](#)
- Tang, S. The key Technology of Lotus Recognition Based on Machine Vision. Master’s Thesis, Jiangsu University, Zhenjiang, China, 2016.
- Septiari, A.; Sunyoto, A.; Hamdani, H.; Kasim, A.A.; Utaminigrum, F.; Hatta, H.R. Machine vision or the maturity classification of oil palm fresh fruit bunches based on color and texture features. *Sci. Hortic.* **2021**, *286*, 110245. [\[CrossRef\]](#)
- Tu, S.; Xue, Y.; Zheng, C.; Qi, Y.; Wan, H.; Mao, L. Detection of passion fruits and maturity classification using Red-Green-Blue Depth images. *Biosyst. Eng.* **2018**, *175*, 156–167. [\[CrossRef\]](#)
- Mima, F.S.; Galibb, S.M.; Hasanc, M.F.; Jerina, S.A. Automatic detection of mango ripening stages—An application of information technology to botany. *Sci. Hortic.* **2018**, *237*, 156–163. [\[CrossRef\]](#)
- Li, H.; Lee, W.S.; Wang, K. Identifying blueberry fruit of different growth stages using natural outdoor color images. *Comput. Electron. Agric.* **2014**, *106*, 91–101. [\[CrossRef\]](#)
- Borrenpohl, D.; Karkee, M. Automated pruning decisions in dormant sweet cherry canopies using instance segmentation. *Comput. Electron. Agric.* **2023**, *207*, 107716. [\[CrossRef\]](#)
- Hameed, K.; Chai, D.; Rassau, A. Score-based mask edge improvement of Mask-RCNN for segmentation of fruit and vegetables. *Expert Syst. Appl.* **2022**, *190*, 116205. [\[CrossRef\]](#)
- Jiang, S.; Liu, Z.; Hua, J.; Zhang, Z.; Zhao, S.; Xie, F.; Ao, J.; Wei, Y.; Lu, J.; Li, Z.; et al. A Real-Time Detection and Maturity Classification Method for Loofah. *Agronomy* **2023**, *13*, 2144. [\[CrossRef\]](#)
- Tang, Y.; Zhou, H.; Wang, H.; Zhang, Y. Fruit detection and positioning technology for a *Camellia oleifera* C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. *Expert Syst. Appl.* **2023**, *211*, 118673. [\[CrossRef\]](#)
- Wang, X.; Kang, H.; Zhou, H.; Au, W.; Chen, C. Geometry-aware fruit grasping estimation for robotic harvesting in apple orchards. *Comput. Electron. Agric.* **2022**, *193*, 106716. [\[CrossRef\]](#)
- Wang, D.; He, D. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [\[CrossRef\]](#)
- He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
- Isaac, P.B.; Diego, M.S.; Manuel, E.G.; Estefanía, C.A. A fast and accurate deep learning method for strawberry instance segmentation. *Comput. Electron. Agric.* **2020**, *178*, 105736.
- Xu, P.; Fang, N.; Liu, N.; Lin, F.; Yang, S.; Ning, J. Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Comput. Electron. Agric.* **2022**, *197*, 106991. [\[CrossRef\]](#)
- Zheng, C.; Chen, P.; Pang, J.; Yang, X.; Chen, C.; Tu, S.; Xue, Y. A mango picking vision algorithm on instance segmentation and key point detection from RGB images in an open orchard. *Biosyst. Eng.* **2021**, *206*, 32–54. [\[CrossRef\]](#)
- Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [\[CrossRef\]](#)
- Kang, H.; Chen, C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. *Comput. Electron. Agric.* **2020**, *171*, 105302. [\[CrossRef\]](#)

20. Kang, H.; Chen, C. Fruit detection and segmentation for apple harvesting using visual sensor in orchards. *Sensors* **2019**, *19*, 4599. [[CrossRef](#)]
21. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
22. Ultralytics. YOLOv8. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 14 June 2024).
23. Wang, H.; Li, Y.; Dang, L.M.; Moon, H. An efficient attention model for instance segmentation network in pest monitoring. *Comput. Electron. Agric.* **2022**, *195*, 106853. [[CrossRef](#)]
24. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
25. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
26. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruit based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [[CrossRef](#)]
27. Ji, W.; Pan, Y.; Xu, B.; Wang, J. A real-time apple targets detection method for picking robot based on ShufflenetV2-YOLOX. *Agriculture* **2022**, *12*, 856. [[CrossRef](#)]
28. Lu, A.; Ma, L.; Cui, H.; Liu, J.; Ma, Q. Instance segmentation of lotus pods and stalks in unstructured planting environment based on improved YOLOv5. *Agriculture* **2023**, *13*, 1568. [[CrossRef](#)]
29. Lu, A.; Liu, J.; Cui, H.; Ma, L.; Ma, Q. MLP-YOLOv5: A Lightweight Multi-Scale Identification Model for Lotus Pods with Scale Variation. *Agriculture* **2023**, *14*, 30. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.