

Article

An Improved YOLOv7-Tiny Method for the Segmentation of Images of Vegetable Fields

Shouwei Wang ^{1,2,3}, Lijian Yao ^{1,2,3,*}, Lijun Xu ^{1,2,3} , Dong Hu ^{1,2,3} , Jiawei Zhou ^{1,2,3} and Yexin Chen ^{1,2,3}

¹ College of Optical, Mechanical and Electrical Engineering, Zhejiang A&F University, Hangzhou 311300, China; w@stu.zafu.edu.cn (S.W.); zoelijun@zafu.edu.cn (L.X.); 20180047@zafu.edu.cn (D.H.); 2022612021033@stu.zafu.edu.cn (J.Z.); 2022612021001@stu.zafu.edu.cn (Y.C.)

² National Engineering Technology Research Center of State Forestry and Grassland Administration on Forestry and Grassland Machinery for Hilly and Mountainous Areas, Hangzhou 311300, China

³ Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in South-Eastern China (Co-Construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Hangzhou 311300, China

* Correspondence: ljyao@zafu.edu.cn

Abstract: In response to the limitations of existing methods in differentiating between vegetables and all types of weeds in farmlands, a new image segmentation method is proposed based on the improved YOLOv7-tiny. Building on the original YOLOv7-tiny framework, we replace the CIOU loss function with the WIOU loss function, substitute the Leaky ReLU loss function with the SiLU activation function, introduce the SimAM attention mechanism in the neck network, and integrate the PConv convolution module into the backbone network. The improved YOLOv7-tiny is used for vegetable target detection, while the *ExG* index, in combination with the OTSU method, is utilized to obtain a foreground image that includes both vegetables and weeds. By integrating the vegetable detection results with the foreground image, a vegetable distribution map is generated. Subsequently, by excluding the vegetable targets from the foreground image using the vegetable distribution map, a single weed target is obtained, thereby achieving accurate segmentation between vegetables and weeds. The experimental results show that the improved YOLOv7-tiny achieves an average precision of 96.5% for vegetable detection, with a frame rate of 89.3 fps, Params of 8.2 M, and FLOPs of 10.9 G, surpassing the original YOLOv7-tiny in both detection accuracy and speed. The image segmentation algorithm achieves a *mIoU* of 84.8% and an *mPA* of 97.8%. This method can effectively segment vegetables and a variety of weeds, reduce the complexity of segmentation with good feasibility, and provide a reference for the development of intelligent plant protection robots.

Keywords: vegetable detection; improved YOLOv7-tiny; image segmentation; *ExG* index



Citation: Wang, S.; Yao, L.; Xu, L.; Hu, D.; Zhou, J.; Chen, Y. An Improved YOLOv7-Tiny Method for the Segmentation of Images of Vegetable Fields. *Agriculture* **2024**, *14*, 856. <https://doi.org/10.3390/agriculture14060856>

Academic Editor: Maciej Zaborowicz

Received: 9 May 2024

Revised: 27 May 2024

Accepted: 28 May 2024

Published: 29 May 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

China is a major producer and consumer of vegetables [1], and vegetables have become the second largest crop in China after grain, ranking first in the world in terms of planting area and production. Accurate and precise fertilization and spraying are crucial for promoting vegetable growth and preventing and controlling weed infestation [2–4]. The traditional way of applying fertilizers and sprays with sloppy amounts of inputs is prone to a waste of agricultural materials and damage to the ecological environment [5]. Therefore, it has become a recent research hotspot to quickly and accurately distinguish vegetables from weeds in fields and develop corresponding precision-to-target fertilizer and drug application strategies [6].

Techniques such as clustering, enhancement, segmentation, and morphological operations are used to highlight features [7–10], and then, features such as the shape, texture, and color of crops and weeds are extracted through edge detection, texture analysis, and color histograms. By combining these features with classifiers such as Support Vector

Machines (SVMs) or Artificial Neural Networks, a distinction between crops and weeds is achieved. However, the above methods rely on manual selection and extraction of features, which are more effective under ideal experimental environments, are susceptible to environmental factors such as background noise and light changes, and have poor generalization ability [11]. In recent years, deep learning has demonstrated tremendous potential in the agricultural sector due to its powerful feature extraction capabilities and excellent generalization abilities [12–15]. As cited in references [16–19], it has been widely applied in the detection and segmentation of crops and weeds in complex farm environments. Zou et al. [20] proposed an improved U-net model for wheat and weed segmenting in field images, achieving a segmentation *mIoU* of 88.98% and an average frame rate of 52 fps on embedded devices. Wu et al. [21] addressed the issue of small-target weed detection by proposing a model based on an improved YOLOv4, which effectively improved the detection accuracy of small target weeds. Wang et al. [22] aimed to improve the detection accuracy of nightshade seedlings with a YOLO-CBAM model, outperforming the original YOLOv5 with a precision of 94.65% and a recall of 90.17%. Yang et al. [23] proposed a new neural network architecture to improve the accuracy of detecting broad-leaf weeds in alfalfa. This architecture integrates the ResNet-101 backbone with image classification and segmentation modules and shows better detection performance for broad-leaf weeds in alfalfa compared to other models. Sahin et al. [24] used multispectral imaging and a CRF-enhanced U-Net model to segment weeds and crops, achieving a *mIoU* of 88.3% on a sunflower dataset, providing a feasible method for early weed detection. Cui et al. [25] proposed a semantic segmentation network, RDS Unet, based on corn seedling fields built upon an improved U-net network. This network accurately recognizes weeds even under complex environmental conditions. The *mIoU*, precision, and recall of this network are 82.36%, 91.36%, and 89.45%, respectively.

All of the above methods employ deep convolutional neural networks to automatically extract image features for the direct detection or segmentation of crops and weeds. Deep learning models require a large number of sample images for training. Thus, the main limitation of this approach is the significant time and labor cost required to construct a target detection or semantic segmentation dataset covering multiple weed categories. In particular, the construction of a semantic segmentation dataset necessitates tedious per-pixel labeling of sample images for model training. Therefore, this method is often limited to recognizing only a single type or a limited number of weed species, restricting its wide application in practical agricultural production environments.

In order to address the limitations of the aforementioned method in distinguishing vegetables and different types of weeds, this paper proposes an image segmentation method based on an improved YOLOv7-tiny. By detecting vegetable targets to indirectly segment vegetables and weeds, the complexity of segmentation can be effectively reduced.

2. Materials and Methods

2.1. Materials

This paper used images of growing pak choi and its accompanying weeds as research objects. The sample images were collected from the vegetable experimental base of Zhejiang A&F University from September to December 2023, including various weather conditions such as sunny, overcast, and cloudy, and from experimental fields with different environmental conditions. The images were captured using an iPhone 13 Pro (Apple Inc., Apple 1 Infinite Loop, Cupertino, CA, USA), with the shooting angle perpendicular to the ground at a height from the ground of approximately 65 cm, totaling 580 images with a resolution of 3840×3840 pixels in JPEG format. Based on practical application scenarios, the sample images were categorized into four cases: weather changes, dense weed growth, similar weed companions, and different weed companions. In order to save training time, the images were cropped to 640×640 pixels. To increase the number and diversity of samples and avoid model overfitting during the training process, operations such as brightness change, adding random noise, and mirroring were applied to augment the sample images. The

number of sample images after augmentation were increased to 2320. The LabelImg1.8.6 software annotated the pak choi targets and generated XML label files. Finally, the sample images were divided into a training set (1856 images), a validation set (232 images), and a test set (232 images) according to the ratio of 8:1:1. Some of the sample images are shown in Figure 1.

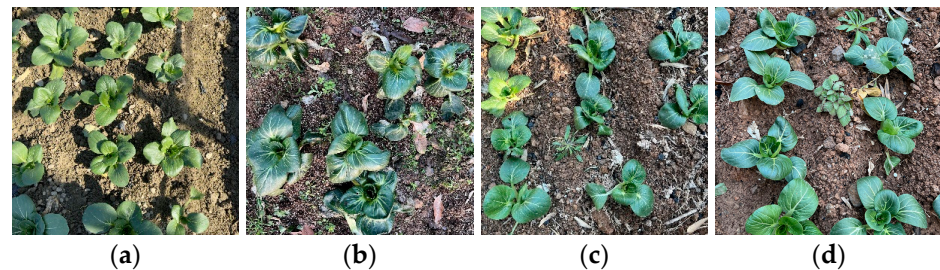


Figure 1. Pak choi sample images. (a) Impact of weather; (b) dense weed growth; (c) similar weed companions; (d) different weed companions.

2.2. Methods

This method is primarily divided into the improved YOLOv7-tiny and the image segmentation algorithm. Firstly, the original YOLOv7-tiny is improved by replacing the loss function with WIoU, introducing the SimAM attention mechanism, switching the activation function to SiLU, and incorporating the PConv convolution module. The improved YOLOv7-tiny is then trained with both the training and validation sets. Secondly, the trained model is used to detect a single pak choi target in the test set, while the foreground map containing both cabbage and weeds is obtained using the ExG index and OTSU method. A pak choi distribution map is generated from the pak choi detection results and the foreground map. Based on this map, pak choi targets are excluded from the foreground map, allowing the weed targets to be obtained individually. Ultimately, precise segmentation of pak choi and weeds is achieved, as illustrated in Figure 2.

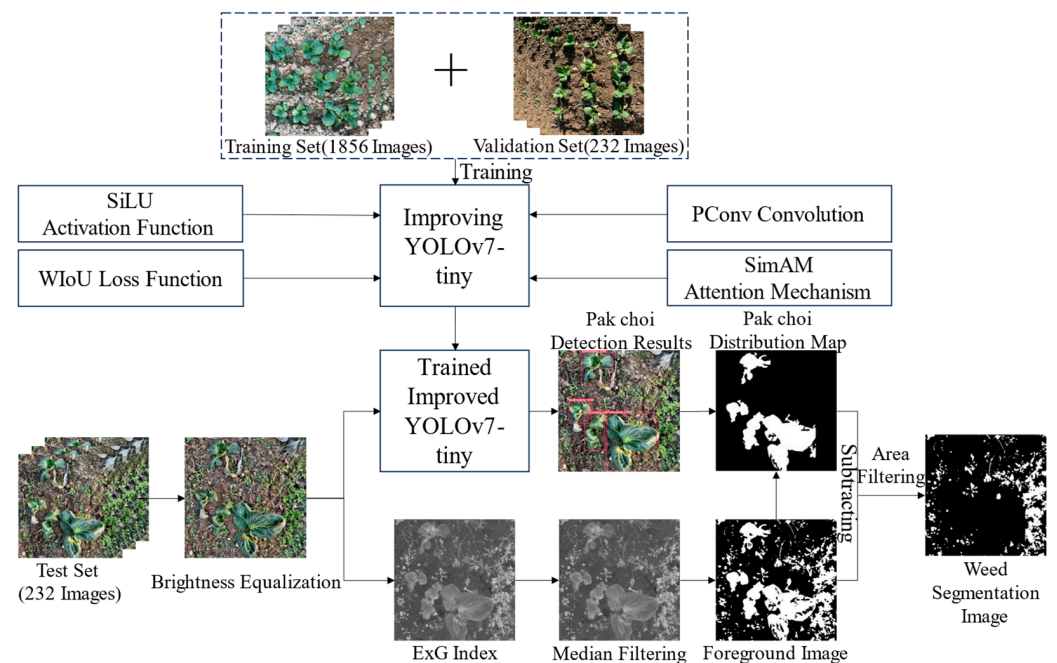


Figure 2. Method flowchart.

2.2.1. Improving YOLOv7-Tiny

YOLOv7-tiny is a compact model of the YOLOv7 series developed for mobile devices [26]. Given the complexity of the open-field vegetable farming environment and weed species diversity, higher requirements are placed on the target detection model. Therefore, to improve the robustness and generalization ability of the model and further enhance the detection accuracy for pak choi, the original YOLOv7-tiny is improved from four aspects. The improved YOLOv7-tiny is divided into four parts: input, backbone, neck, and head, as shown in Figure 3.

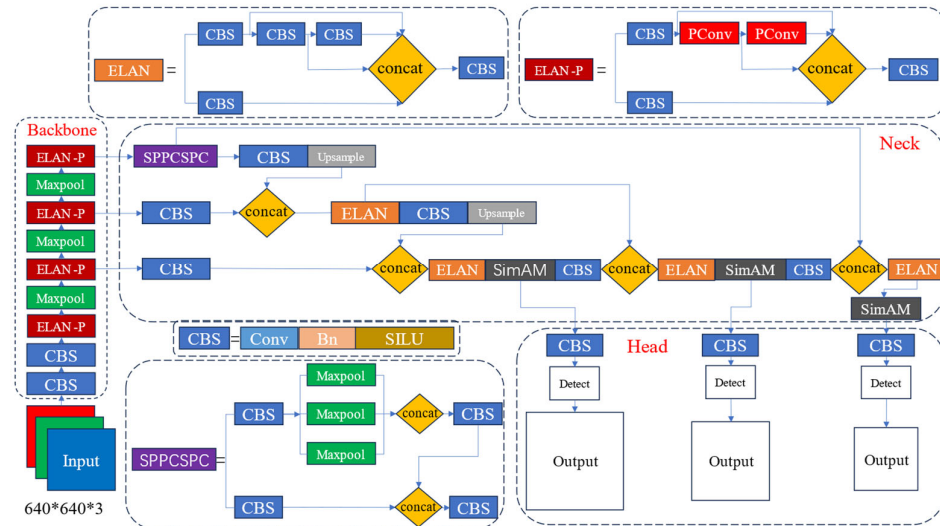


Figure 3. Structural diagram of improved YOLOv7-tiny.

1. WIoUv3 Loss Function

The original YOLOv7-tiny utilizes the CIoU loss function [27], which, building on the distance intersection over union (DIOU) loss [28], accounts for the bounding box's overlapping area, center distance, and aspect ratio. Since the SCT dataset comes from autonomous collection and labeling, the training set inevitably contains low-quality samples (such as inaccurate labeling). Using the CIoU loss function, which considers factors such as center distance and aspect ratio, might excessively penalize low-quality samples, affecting the model's generalization ability. While WIoUv1 with a dual-layer attention mechanism mitigates this issue, its advanced version, WIoUv3 [29], featuring a dynamic non-monotonicity FM, helps to improve the convergence rate of the model further, reducing negative impacts during training and improving the trained model's robustness. Therefore, this paper adopts WIoUv3 as the loss function to balance the impact of low and high-quality samples, enhance anchor box quality, and thereby improve detection and localization accuracy. The formula for WIoUv3 is as follows:

$$\begin{cases} L_{WIoUv3} = r \times R_{WIoU} \times L_{IoU} \\ R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \\ \beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \\ r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \end{cases} \quad (1)$$

In Equation (1), (x_{gt}, y_{gt}) and (x, y) are the coordinates of the center points of the ground actual box and the predicted box, respectively; W_g and H_g are the width and height of the smallest region containing both the ground actual box and the predicted box; R_{WIoU} is the penalty term; β represents the outliers, used to measure the quality of the anchor boxes, with lower outliers indicating higher box quality; r is the non-monotonic focusing coefficient; and α and δ are hyperparameters.

2. SimAM Parameter-Free Attention Mechanism

In open-field vegetable fields, the lighting conditions are complex, and pak choi overlaps with weeds, obstructing the view. The parameter-free attention mechanism SimAM [30] can adaptively emphasize the target features of pak choi and suppress irrelevant background features (e.g., weeds, soil) without increasing model complexity. Unlike traditional channel and spatial attention mechanisms, SimAM possesses total three-dimensional weights to better refine the extracted pak choi features. SimAM is introduced into the neck network to optimize the features extracted from the backbone network, enhance the model's anti-interference ability, and help it identify and locate pak choi more accurately in complex field environments. The structure of SimAM is shown in Figure 4.

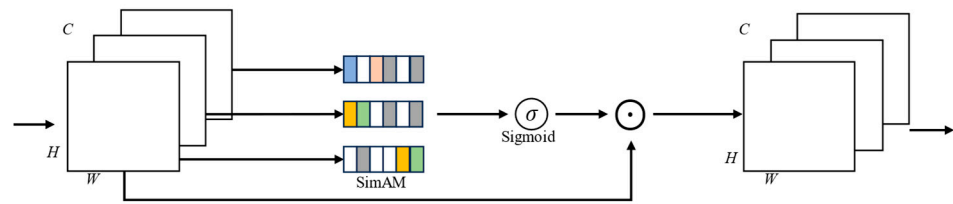


Figure 4. SimAM structure diagram.

For the feature map $X \in \mathbb{R}^{C \times W \times H}$, the SimAM attention mechanism uses the minimum energy e_t^* to measure the importance of the target neuron t , with e_t^* defined as follows:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (2)$$

In Equation (2), λ is the regularization term; t represents the target neuron of the input feature map on a single channel, $\hat{\mu}$ is the mean of all neurons on a single channel; $\hat{\sigma}^2$ is the variance of all neurons on a single channel.

The lower the minimum energy e_t^* , the more different the target neuron t in the pak choi features is from the surrounding neurons, and the higher its importance. The expressions for $\hat{\sigma}^2$ and $\hat{\mu}$ are as follows:

$$\begin{cases} \hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^M (x_i - \hat{\mu})^2 \\ \hat{\mu} = \frac{1}{M} \sum_{i=1}^M x_i \end{cases} \quad (3)$$

In Equation (3), x_i denotes the number of neurons, corresponding to all neurons in the dimension of the target neuron.

Finally, according to the definition of the attention mechanism, the output features $\tilde{X}$ are obtained by suppressing the oversized values through the Sigmoid function as follows:

$$\tilde{X} = \text{Sigmoid}\left(\frac{1}{e_t^*}\right) \odot X \quad (4)$$

3. SiLU Activation Function

The original YOLOv7-tiny uses the Leaky ReLU activation function, which remains linear for non-negative values and prevents gradient vanishing in the negative part by a slight slope. However, Leaky ReLU is not sufficiently smooth near the zero point, which affects the gradient update and feature extraction. Considering the diversity of pak choi target features under different lighting, shading, and weed interference, the model requires enhanced feature extraction and representation capabilities. Therefore, Leaky ReLU is replaced by the SiLU activation function [31], utilizing its smooth and non-monotonic characteristics, as well as the feature of no upper bound and a lower bound, to strengthen

the model's regularization effect and further enhance its ability to extract and learn target features. The expression for the SiLU activation function is as follows:

$$\text{SiLU}(x) = \frac{x}{1 + e^{-x}} \quad (5)$$

4. Design of ELAN-P Module Based on Pconv

Partial convolution (PConv) [32] effectively extracts spatial features by reducing the number of redundant computations and memory accesses. As illustrated in Figure 5, it exploits redundancy between channels to perform convolution operations only on a subset of the input channels, leaving the other channels unchanged. Therefore, PConv outperforms conventional and depth-wise convolutions regarding the number of floating-point operations and computational speed, efficiently expressing the model with fewer parameters and thus reducing computational load and memory consumption.

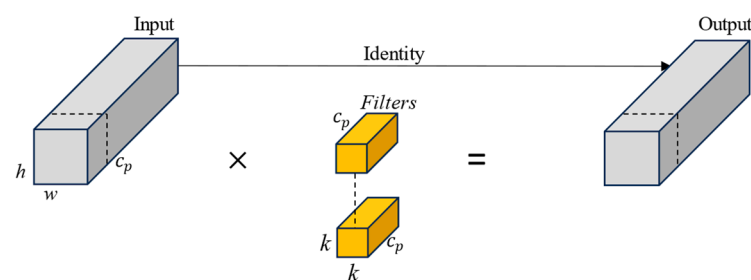


Figure 5. PConv principle diagram.

The FLOPs of PConv are " $h \times w \times k^2 \times c_p^2$ ", which is 1/16 of that for conventional convolution. Moreover, the memory access of PConv is " $h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p$ " which is 1/4 of that for conventional convolution. The FLOPs of PConv are lower than that of the conventional convolution but higher than that of the depth-wise convolution (DWConv). Since this study focuses solely on detecting the single target of pak choi and does not require complex depth-wise convolution operations for feature extraction, the standard convolutions with a kernel size of 3×3 and a step size of 1 are replaced with PConv. As shown in Figure 6, Figure 6a is the ELAN module of the original YOLOv7-tiny, and Figure 6b is the improved ELAN-P module. Replacing some ELAN modules in the backbone network with ELAN-P modules not only lightens the network but also aggregates information to enhance the generation of pak choi feature maps, aiding in the performance improvement in YOLOv7-tiny.

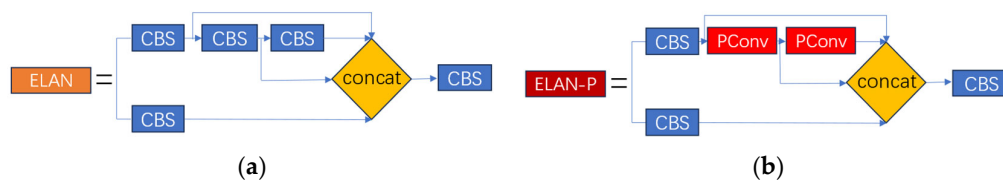


Figure 6. Design of ELAN-P module. (a) ELAN module; (b) ELAN-P module.

2.2.2. Image Segmentation Algorithm

1. Image Brightness Equalization

The image brightness equalization method, designed to mitigate the impact of illumination on segmentation quality, follows a systematic approach. It starts by converting the image from the RGB to HSV color space and extracting the brightness channel V. The average brightness of the V channel is then calculated. The V channel is divided into 32 blocks, with the average brightness of each block computed. Finally, the difference between the average brightness of each block and the overall average brightness is calculated, ensuring a reliable and consistent process.

Principle: assuming a 640×640 pixel image has a grayscale level of $(0, 1, 2, \dots, L)$, the average brightness of this image is as follows:

$$Lum_{av} = \frac{\sum_{i=0}^{640-1} \sum_{j=0}^{640-1} p(i, j)}{640 \times 640} \quad (6)$$

The average brightness of each sub-block is as follows:

$$Lum_{av_{bm}} = \frac{\sum_{i=0}^{32-1} \sum_{j=0}^{32-1} p(i, j)}{32 \times 32} \quad (7)$$

In Equations (6) and (7), $p(i, j)$ represents the brightness value of the image at coordinates (i, j) .

Then, the difference between the average brightness of the image sub-block and the overall average brightness is as follows:

$$\Delta_{lum} = Lum_{av_{bm}} - Lum_{av} \quad (8)$$

From Equation (8), it is known that if it is positive, then the image sub-block is of high brightness; otherwise, it is of low brightness. The brightness of image sub-blocks with a positive difference is weakened, while that of sub-blocks with a negative difference is enhanced. The matrix of the image sub-blocks undergoes bilinear interpolation according to the image segmentation method, and it is expanded to the size of the entire original image. The expanded matrix subtracts the original V channel image to obtain the equalized V channel image. The equalized V channel image is then merged with the H and S channels of the original image and converted back to the RGB color space, resulting in the brightness-equalized image.

The brightness variance in the original image is 3759, while that of the brightness-equalized image is 2621. Comparing the variances of the two images, it is evident that the original image has noticeably too dark and too bright areas due to lighting, while the light-equalized image has a more uniform overall brightness distribution, as shown in Figure 7.

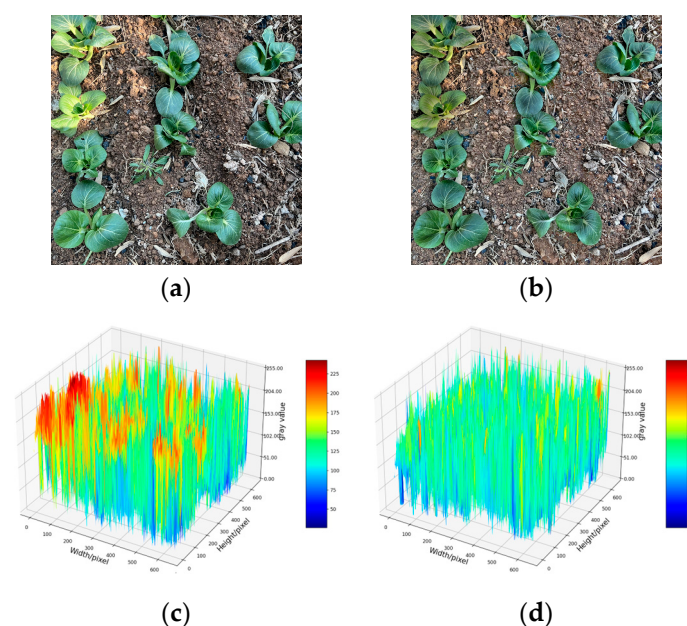


Figure 7. Illustration of brightness equalization effect. (a) Original image; (b) brightness-equalized image; (c) grayscale 3D plot of original image; (d) brightness-equalized grayscale 3D plot.

2. Pak Choi Detection and the ExG Index

Inputting the brightness-equalized image into the improved YOLOv7-tiny outputs the location information of the pak choi detection boxes. Simultaneously, utilizing the color differences among pak choi, weeds, and soil, the weight of the green channel is calculated and enhanced, i.e., the ExG index [33], to obtain a grayscale image highlighting the green pixels, as shown in Figure 8. The formula for calculating the ExG index is as follows:

$$ExG = \begin{cases} 0, & 2g < r + b \\ 2g - r - b, & otherwise \end{cases} \quad (9)$$



Figure 8. Pak choi detection and ExG index illustration. (a) Pak choi detection results; (b) ExG index grayscale image.

In Equation (9), r , g , and b represent the normalized RGB values in the RGB color space, and the calculation formula is as follows:

$$r = \frac{R}{R + G + B}, g = \frac{G}{R + G + B}, b = \frac{B}{R + G + B} \quad (10)$$

3. Binarized Foreground Image

The grayscale image obtained by using $E \times G$ feature computation ultimately retains the pak choi and weeds and effectively suppresses the soil background. Median filtering arranges all pixel grayscale values within a 7×7 window from most significant to most minor. It replaces the grayscale value at point (i, j) with the median grayscale value within the window, effectively removing environmental background noise. After noise reduction through median filtering, the grayscale image uses the OTSU thresholding method to determine the optimal segmentation threshold, dividing the image into foreground and background. Subsequently, using the closing operation to eliminate noise points and fill voids, a foreground image containing pak choi and weeds is obtained, as shown in Figure 9.

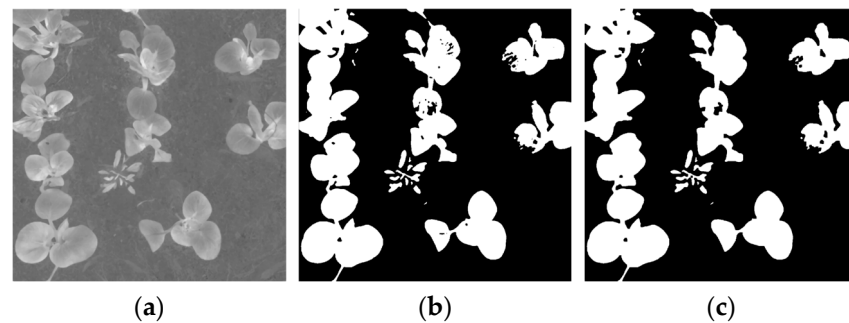


Figure 9. Binarized foreground image: (a) 7×7 median filtering; (b) foreground image; (c) closing operation.

4. Pak Choi and Weed Segmentation

Based on the detection results from the improved YOLOv7-tiny, a pak choi distribution image is created by preserving the most significant area contour within the detection boxes. Subtracting this image from the pak choi and weed foreground image yields a binary image containing only weeds. Finally, the final weed segmentation image is obtained by applying area filtering to process noise in the image, as shown in Figure 10.

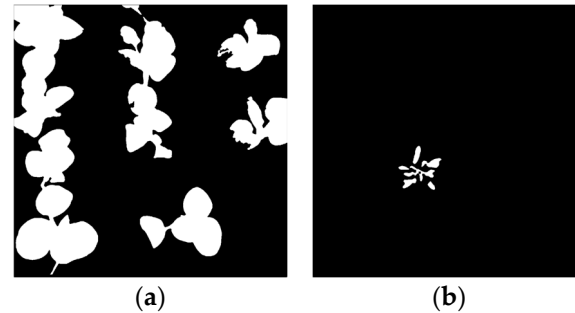


Figure 10. Pak choi distribution map and weed segmentation image. (a) Pak choi distribution map; (b) weed segmentation image.

2.3. Experimental Environment

This study was conducted using Python 3.8.17 and the Pytorch 1.13.0 framework, and the experiments were conducted on an NVIDIA GeForce RTX 3080 (Santa Clara, CA, USA). The parameter settings are shown in Table 1.

Table 1. Experimental parameter settings.

Parameter	Value
Input Image Size/Pixel	640 × 640
Batch Size	16
Momentum	0.937
Initial Learning Rate	0.01
Weight Decay	0.0005
Warmup	3
Confidence Threshold	0.5
Non-Maximum Suppression IoU Threshold	0.5
Epoch	300

2.4. Evaluation Metrics

2.4.1. Evaluation Metrics for Improved YOLOv7-Tiny

The evaluation metrics for the improved YOLOv7-tiny include precision (P), recall (R), average precision (AP), Params, FLOPs, and frame rate (FPS). The formulas for calculating precision and recall are as follows:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \end{cases} \quad (11)$$

In Equation (11), TP (true positives) is the number of correctly identified positive instances; FP (false positives) is the number of incorrectly identified positive instances; and FN (false negatives) is the number of incorrectly rejected negative instances.

Average precision (AP) is a key metric that measures the overall detection performance of the model. It is the average of precision at different recall levels. A higher AP value indicates better detection performance. The calculation formula is as follows:

$$AP = \int_0^1 P(R) dR \quad (12)$$

Params, which refers to the total number of trainable parameters in the model, measured in millions (M), and FLOPs, representing the number of floating-point operations required for the model to perform inference, measured in billions (G), are two technical terms you should be familiar with in our model.

2.4.2. Evaluation Metrics for Image Segmentation Algorithms

The evaluation metrics for image segmentation algorithms include pixel accuracy (PA), mean pixel accuracy (mPA), intersection over union (IoU), mean intersection over union (mIoU), and frames per second (FPS).

The formulas for calculating mIoU and mPA are as follows:

$$\left\{ \begin{array}{l} mIoU = \frac{1}{k+1} \frac{\sum_{i=0}^k \frac{P_{ij}}{\sum_{j=0}^k P_{ij} + \sum_{j=0}^k P_{ji} - P_{ii}}}{\sum_{j=0}^k P_{ij}} \\ mPA = \frac{1}{k+1} \frac{\sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij}}}{\sum_{j=0}^k P_{ij}} \end{array} \right. \quad (13)$$

In Equation (13), $k + 1$ represents the number of classes. P_{ij} denotes the number of pixels of class i predicted as class j ; thus, P_{ii} is true positives, P_{ij} is false negatives, and P_{ji} is false positives.

3. Results and Discussion

3.1. Experiment and Analysis of Improved YOLOv7-Tiny

3.1.1. Different Loss Functions

The loss function was replaced with commonly targeted SIoU [34], Focal EIou [35], and WIoUv3 to verify the impact of different loss functions on the evaluation metrics. The experimental results are shown in Table 2; the loss curves for different loss functions on the training set as the training steps increase are depicted in Figure 11.

Table 2. Evaluation metrics under different loss functions.

Loss Function	Precision/%	Recall/%	AP/%
CIoU	92.4	92.7	93.4
SIoU	91.4	92.9	93.1
Focal EIou	91.1	93.1	92.3
WIoUv3	92.8	92.5	93.7

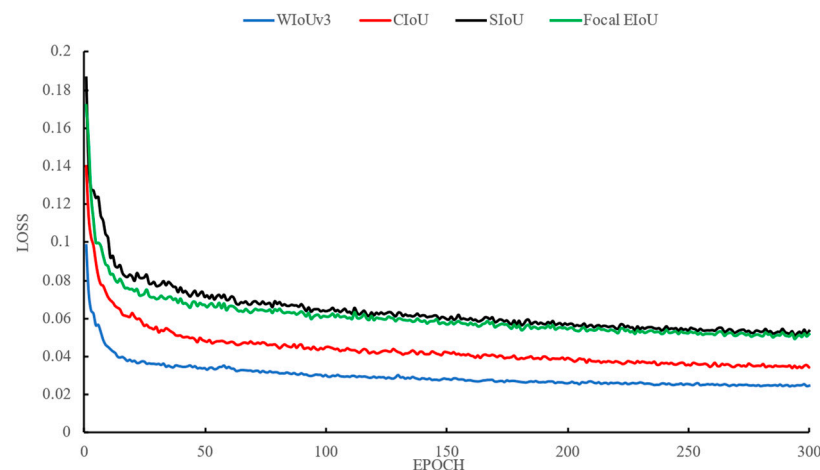


Figure 11. Loss curves for different loss Functions over training steps.

As shown in Figure 11, the model using the WIoUv3 loss function starts with a lower initial loss during training, has a high learning rate in the first ten iterations, and rapidly decreases in loss, demonstrating quick fitting capability. From the 10th to the 263rd cycle, the loss decreases slowly and stabilizes after 270 cycles, fluctuating around 0.024. This indicates that this study's model trained with WIoUv3 exhibits more robust performance and better robustness.

As indicated by Table 2, the models trained with the SIoU and Focal EIoU loss functions have lower precision, recall, and *AP* values compared to those trained with the CIoU loss function, suggesting that SIoU and Focal EIoU are not suitable for the dataset used in this study. Compared to CIoU, the model trained with the WIoUv3 loss function shows a 0.4% increase in precision, a slight 0.2% decrease in recall, and a 0.3% increase in *AP*. The result demonstrates that the WIoUv3-trained model used in this study has superior performance and better robustness.

3.1.2. Incorporating SimAM at Different Locations

To analyze the impact of incorporating the SimAM attention mechanism at different positions on detection performance, SimAM was introduced separately into the backbone and neck networks of the original YOLOv7-tiny, and the experimental results were analyzed. The results are shown in Table 3.

Table 3. Evaluation metrics for attention mechanism embedded at different positions.

Model	Precision/%	Recall/%	<i>AP</i> /%	Params/M	FLOPs/G
Original YOLOv7-tiny	92.4	92.7	93.4	11.6	13.2
+SimAM(Backbone)	91.2	90.1	92.3	7.3	9.1
+SimAM(Neck)	93.1	94.8	94.2	11.0	13.6

As indicated by Table 3, integrating SimAM into the backbone network reduces the model's Params and FLOPs but lowers the *AP*. However, incorporating SimAM into the neck network, though slightly increasing the FLOPs (+0.4 G), improves the *AP* by 0.8% and reduces Params by 0.6 M. SimAM is less effective in the backbone network than in the neck network, possibly because the pak choi features extracted in the backbone lack distinct discriminability, making it difficult for SimAM to accurately identify the most critical feature channels for pak choi in the backbone. On the other hand, the neck network processes the extracted pak choi features further; at this stage, the input features are more abundant and distinctive, with fewer feature channels than in the backbone. In this context, introducing SimAM can more effectively capture information about pak choi features, further enhancing the expression of pak choi characteristics.

3.1.3. Ablation Experiments

Experiments with five different configurations were conducted to visually analyze the enhancements in the improved YOLOv7-tiny compared to the original model, and the various evaluation metrics were compared. The results of the ablation study are shown in Table 4.

Table 4. Ablation study results.

Combination Number	Structural Configuration						Evaluation Metrics			
	Leaky ReLU	CIoU	SILU	WIoUv3	SimAM	PConv	<i>AP</i> /%	FPS	Params/M	FLOPs/G
1	✓	✓	×	×	×	×	93.4	79.4	11.6	13.2
2	×	✓	✓	×	×	×	94.8	85.2	11.6	13.2
3	×	×	✓	✓	×	×	95.3	89.3	11.6	13.2
4	×	×	✓	✓	✓	×	96.4	89.3	11.0	13.6
5	×	×	✓	✓	✓	✓	96.5	89.3	8.2	10.9

Configuration 1 is the original YOLOv7-tiny, with an AP of 93.4%, 11.6 M Params, 13.2 G FLOPs, and 79.4 fps. Configuration 2, based on Configuration 1, replaces the Leaky ReLU activation function with SILU, improving the AP to 94.8% (+1.4%) and increasing the fps to 85.2 (+5.8). Configuration 3, building on Configuration 2, replaces CIOU with WIOUv3, raising the AP to 95.3% (+0.5%). Configuration 4, based on Configuration 3, introduces SimAM into the neck network, increasing the AP to 96.4% (+1.1%), reducing Params to 11.0 M (−5.1%), and increasing FLOPs to 13.6 G (+2.9%). Configuration 5, based on Configuration 4, replaces part of the backbone network’s ELAN modules with ELAN-P modules (based on PConv), with no significant change in AP , a 25% reduction in Params, and a 20% decrease in FLOPs. In summary, Configuration 5 is the final improved YOLOv7-tiny, which, compared to the original YOLOv7-tiny, has increased the AP by 3.1% and the fps by 12%, and reduced Params by 29% and FLOPs by 17%. Overall, these metrics indicate that the improved YOLOv7-tiny has superior performance. Some detection results are shown in Figure 12.

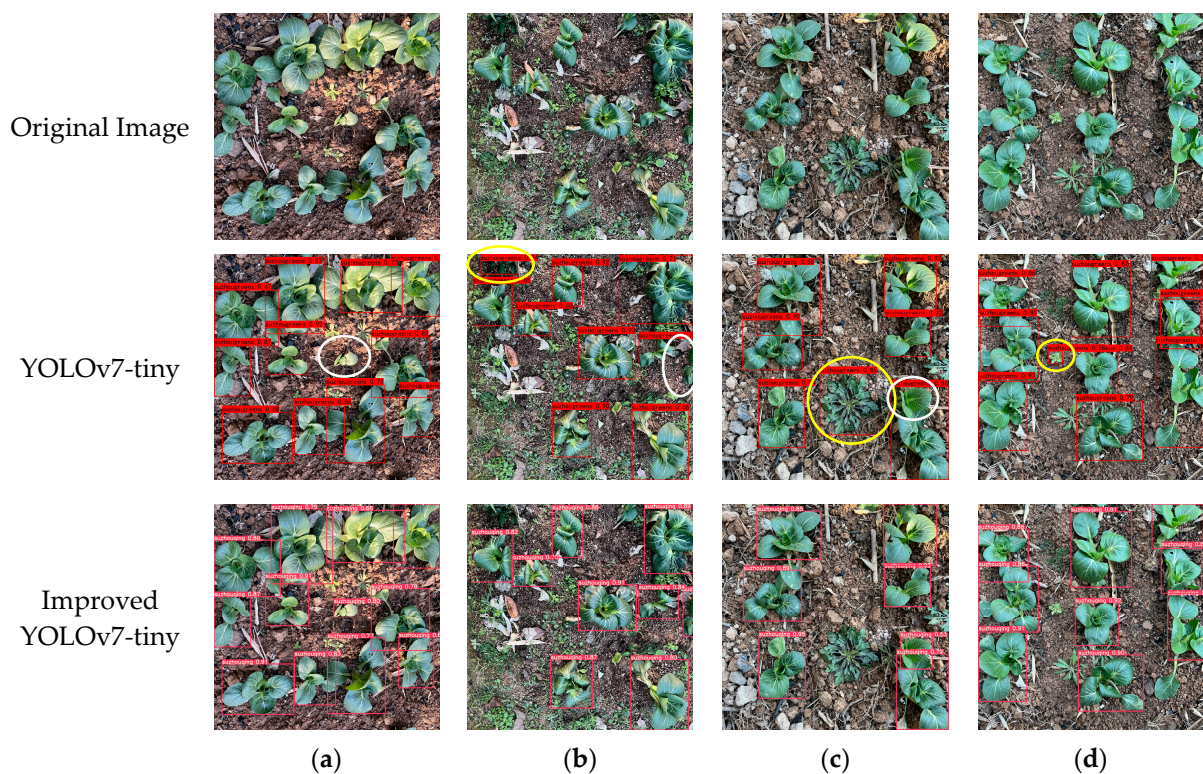


Figure 12. Comparison of pak choi detection results. (a) Impact of weather; (b) dense weed growth; (c) similar accompanying weeds; (d) different accompanying weeds. Note: yellow circles indicate false detections, and white circles indicate missed detections.

Figure 12a shows a complex lighting environment, where some smaller pak choi plants are missed due to poor visibility. In Figure 12b, the weeds are densely grown, and the weeds in the upper left corner of the image are mistakenly identified as pak choi by the original YOLOv7-tiny due to their proximity to the pak choi. In Figure 12c, because the weeds and pak choi are similar in color and shape, the original YOLOv7-tiny incorrectly identifies the weeds as pak choi and recognizes two closely connected pak choi plants in the bottom right corner of the image as one. Figure 12d shows a variety of weeds coexisting, and due to the insufficient interference resistance of the original YOLOv7-tiny, false detections occur.

3.2. Image Segmentation Experiments and Analysis

The method described in Section 2.2.2 of this paper was used to segment the 232 images in the test set, and the quality evaluation metrics of image segmentation are shown in Table 5.

Table 5. Image segmentation experiment results.

Type	<i>IoU</i> /%	<i>PA</i> /%	<i>mIoU</i> /%	<i>mPA</i> /%	FPS
Weeds	76.5	97.2	84.8	97.8	62.5
Pak choi	93.1	98.4			

As shown in Table 5, the *IoU* for weeds is 76.5%, and the *PA* is 97.2%; for pak choi, the *IoU* is 93.1%, and the *PA* is 98.4%. The *mIoU* and *mPA* for both are 84.8% and 97.8%, respectively, with an fps value of 62.5. These results demonstrate that the method achieves high segmentation accuracy and good real-time performance when distinguishing between pak choi and weeds. Some segmentation results are shown in Figure 13.

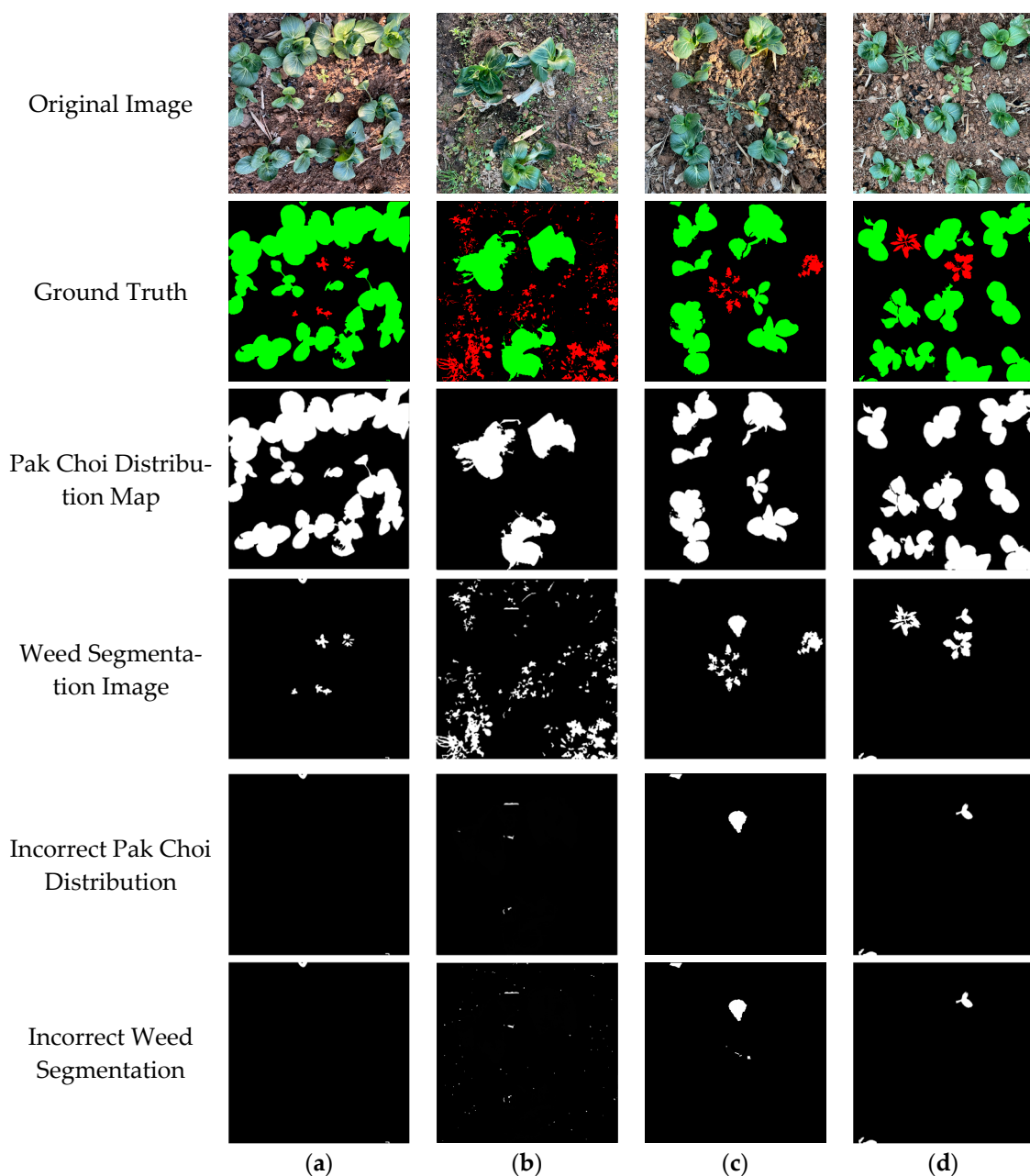


Figure 13. Image segmentation results. (a) Impact of weather; (b) dense weed growth; (c) similar accompanying weeds; (d) different accompanying Weeds.

A detailed analysis of the segmentation results has revealed several key factors affecting segmentation accuracy: The improved YOLOv7-tiny exhibits lower accuracy when detecting pak choi at the image edges and in highly incomplete forms, leading to incorrect segmentation of pak choi, as shown in Figure 13a. Smaller weed pixels are considered noise and are filtered out during area filtering, as seen in Figure 13b,c. Weeds that come into contact with pak choi and are within the detection frame are incorrectly segmented as pak choi pixels. Additionally, non-green debris on the pak choi, such as soil clumps and branches, and detection frames that do not fully cover the pak choi result in incomplete segmented pak choi targets, as depicted in Figure 13d. These findings highlight areas for improvement and offer guidance for enhancing segmentation accuracy in future work.

In comparison to traditional methods that utilize deep learning models to directly segment crops from weeds [20,24,25], the main advantages of the approach proposed in this paper are as follows:

1. Building a semantic segmentation dataset that includes various types of weeds is an extremely cumbersome task. By contrast, the method outlined in this paper only requires the creation of a target detection dataset for crops to train the model, significantly reducing the cost and difficulty of dataset construction.
2. This method segments crops from weeds indirectly by detecting the crops, thereby eliminating the need to segment each type of weed individually. This reduces the complexity of segmentation and enhances the robustness compared to direct segmentation of crops and weeds.

4. Conclusions

1. This paper focuses on pak choi and its accompanying weeds as the subjects of study and proposes an image segmentation method based on an improved YOLOv7-tiny. It detects pak choi and segments it and the weeds, effectively reducing the complexity of segmentation.
2. Building on the original YOLOv7-tiny, the WIoU loss function and SiLU activation function are adopted to replace the existing loss function and activation function, the SimAM attention mechanism is introduced into the neck network, and the PConv convolution module is integrated into the backbone network. Compared to the original YOLOv7-tiny, the improved YOLOv7-tiny has an increased *AP* by 3.1%, increased fps by 12%, reduced Params by 29%, and decreased FLOPs by 17%. These improvements reduce the model's consumption and significantly enhance the detection accuracy and speed for pak choi.
3. The improved YOLOv7-tiny is used to identify individual pak choi targets in farmland, combined with the *ExG* index and OTSU method, to obtain a foreground image containing both pak choi and weeds. A pak choi distribution map is created by combining the target detection results of pak choi with the foreground image. Subsequently, a single weed target is obtained using the pak choi distribution map to remove pak choi targets from the foreground image, achieving precise segmentation between pak choi and weeds. The specific evaluation metrics for image segmentation are an *mIoU* of 85.3%, an *mPA* of 97.8%, and fps of 62.5. These results validate the efficiency of the proposed method in terms of segmentation accuracy and real-time performance.
4. Despite this method showing strong feasibility, there are still some limitations: firstly, the accuracy of the improved YOLOv7-tiny in detecting pak choi at the edges of images needs to be improved; secondly, weeds that touch pak choi and are within the detection frame can be incorrectly identified as pak choi pixels; finally, debris such as branches and soil clumps on the pak choi may result in incomplete segmented pak choi targets. Future research will seek improvement measures for these shortcomings to enhance the segmentation accuracy further.

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

Funding: This study was funded by the Key R&D Program of Zhejiang, grant numbers (2022C02042) and (2023C02053).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

- Han, J.; Luo, Y.; Yang, L.; Liu, X.; Wu, L.; Xu, J. Acidification and salinization of soils with different initial pH under greenhouse vegetable cultivation. *J. Soils Sediments* **2014**, *14*, 1683–1692. [[CrossRef](#)]
- Sharpe, S.M.; Schumann, A.W.; Yu, J.; Boyd, N.S. Vegetation detection and discrimination within vegetable plasticulture row-middles using a convolutional neural network. *Precis. Agric.* **2020**, *21*, 264–277. [[CrossRef](#)]
- Su, W.-H.; Fennimore, S.A.; Slaughter, D.C. Development of a systemic crop signalling system for automated real-time plant care in vegetable crops. *Biosyst. Eng.* **2020**, *193*, 62–74. [[CrossRef](#)]
- Zhang, Y.; Staab, E.S.; Slaughter, D.C.; Giles, D.K.; Downey, D. Automated weed control in organic row crops using hyperspectral species identification and thermal micro-dosing. *Crop Prot.* **2012**, *41*, 96–105. [[CrossRef](#)]
- Franco, C.; Pedersen, S.M.; Papaharalampos, H.; Ørum, J.E. The value of precision for image-based decision support in weed management. *Precis. Agric.* **2017**, *18*, 366–382. [[CrossRef](#)]
- Raja, R.; Nguyen, T.T.; Slaughter, D.C.; Fennimore, S.A. Real-time weed-crop classification and localisation technique for robotic weed control in lettuce. *Biosyst. Eng.* **2020**, *192*, 257–274. [[CrossRef](#)]
- Bakhsipour, A.; Jafari, A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. *Comput. Electron. Agric.* **2018**, *145*, 153–160. [[CrossRef](#)]
- Rehman, T.U.; Zaman, Q.U.; Chang, Y.K.; Schumann, A.W.; Corscadden, K.W. Development and field evaluation of a machine vision based in-season weed detection system for wild blueberry. *Comput. Electron. Agric.* **2019**, *162*, 1–13. [[CrossRef](#)]
- Tang, J.-L.; Chen, X.-Q.; Miao, R.-H.; Wang, D. Weed detection using image processing under different illumination for site-specific areas spraying. *Comput. Electron. Agric.* **2016**, *122*, 103–111. [[CrossRef](#)]
- Wang, A.; Zhang, W.; Wei, X. A review on weed detection using ground-based machine vision and image processing techniques. *Comput. Electron. Agric.* **2019**, *158*, 226–240. [[CrossRef](#)]
- Xu, K.; Shu, L.; Xie, Q.; Song, M.; Zhu, Y.; Cao, W.; Ni, J. Precision weed detection in wheat fields for agriculture 4.0: A survey of enabling technologies, methods, and research challenges. *Comput. Electron. Agric.* **2023**, *212*, 108106. [[CrossRef](#)]
- Albahar, M. A survey on deep learning and its impact on agriculture: Challenges and opportunities. *Agriculture* **2023**, *13*, 540. [[CrossRef](#)]
- Attri, I.; Awasthi, L.K.; Sharma, T.P.; Rathee, P. A review of deep learning techniques used in agriculture. *Ecol. Inform.* **2023**, *77*, 102217. [[CrossRef](#)]
- Li, Y.; Ma, R.; Zhang, R.; Cheng, Y.; Dong, C. A tea buds counting method based on YOLOV5 and Kalman filter tracking algorithm. *Plant Phenomics* **2023**, *5*, 0030. [[CrossRef](#)] [[PubMed](#)]
- Wang, S.; Wu, D.; Zheng, X. TBC-YOLOv7: A refined YOLOv7-based algorithm for tea bud grading detection. *Front. Plant Sci.* **2023**, *14*, 1223410. [[CrossRef](#)] [[PubMed](#)]
- Hasan, A.M.; Soheli, F.; Diepeveen, D.; Laga, H.; Jones, M.G. A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* **2021**, *184*, 106067. [[CrossRef](#)]
- Lottes, P.; Behley, J.; Milioto, A.; Stachniss, C. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2870–2877. [[CrossRef](#)]
- Rakhmatulin, I.; Kamilaris, A.; Andreasen, C. Deep neural networks to detect weeds from crops in agricultural environments in real-time: A review. *Remote Sens.* **2021**, *13*, 4486. [[CrossRef](#)]
- Zhuang, J.; Li, X.; Bagavathiannan, M.; Jin, X.; Yang, J.; Meng, W.; Li, T.; Li, L.; Wang, Y.; Chen, Y. Evaluation of different deep convolutional neural networks for detection of broadleaf weed seedlings in wheat. *Pest Manag. Sci.* **2022**, *78*, 521–529. [[CrossRef](#)]
- Zou, K.; Liao, Q.; Zhang, F.; Che, X.; Zhang, C. A segmentation network for smart weed management in wheat fields. *Comput. Electron. Agric.* **2022**, *202*, 107303. [[CrossRef](#)]
- Wu, H.; Wang, Y.; Zhao, P.; Qian, M. Small-target weed-detection model based on YOLO-V4 with improved backbone and neck structures. *Precis. Agric.* **2023**, *24*, 2149–2170. [[CrossRef](#)]

22. Wang, Q.; Cheng, M.; Huang, S.; Cai, Z.; Zhang, J.; Yuan, H. A deep learning approach incorporating YOLO v5 and attention mechanisms for field real-time detection of the invasive weed *Solanum rostratum* Dunal seedlings. *Comput. Electron. Agric.* **2022**, *199*, 107194. [\[CrossRef\]](#)
23. Yang, J.; Chen, Y.; Yu, J. Convolutional neural network based on the fusion of image classification and segmentation module for weed detection in alfalfa. *Pest Manag. Sci.* **2024**, *80*, 2751–2760. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Sahin, H.M.; Miftahushudur, T.; Grieve, B.; Yin, H. Segmentation of weeds and crops using multispectral imaging and CRF-enhanced U-Net. *Comput. Electron. Agric.* **2023**, *211*, 107956. [\[CrossRef\]](#)
25. Cui, J.; Tan, F.; Bai, N.; Fu, Y. Improving U-net network for semantic segmentation of corns and weeds during corn seedling stage in field. *Front. Plant Sci.* **2024**, *15*, 1344958. [\[CrossRef\]](#)
26. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
27. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [\[CrossRef\]](#)
28. Dong, C.; Duoqian, M. Control distance IoU and control distance IoU loss for better bounding box regression. *Pattern Recognit.* **2023**, *137*, 109256. [\[CrossRef\]](#)
29. Hu, D.; Yu, M.; Wu, X.; Hu, J.; Sheng, Y.; Jiang, Y.; Huang, C.; Zheng, Y. DGW-YOLOv8: A small insulator target detection algorithm based on deformable attention backbone and WIoU loss function. *IET Image Process.* **2024**, *18*, 1096–1108. [\[CrossRef\]](#)
30. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 11863–11874.
31. Elfving, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [\[CrossRef\]](#)
32. Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
33. Zhang, M.; Zhou, J.; Sudduth, K.A.; Kitchen, N.R. Estimation of maize yield and effects of variable-rate nitrogen application using UAV-based RGB imagery. *Biosyst. Eng.* **2020**, *189*, 24–35. [\[CrossRef\]](#)
34. Cui, M.; Duan, Y.; Pan, C.; Wang, J.; Liu, H. Optimization for anchor-free object detection via scale-independent GIoU loss. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [\[CrossRef\]](#)
35. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [\[CrossRef\]](#)

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.