


## Article

# Improving Walnut Images Segmentation Using Modified UNet3+ Algorithm

Jun Tie <sup>1,2</sup> , Weibo Wu <sup>1,3</sup>, Lu Zheng <sup>1,2,\*</sup>, Lifeng Wu <sup>1,3</sup> and Ting Chen <sup>1,3</sup><sup>1</sup> College of Computer Science, South-Central Minzu University, Wuhan 430074, China<sup>2</sup> Hubei Provincial Engineering Research Center of Agricultural Blockchain and Intelligent Management, Wuhan 430074, China<sup>3</sup> Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises, Wuhan 430074, China

\* Correspondence: zhenglu@scuec.edu.cn

**Abstract:** When aiming at the problems such as missed detection or misdetection of recognizing green walnuts in the natural environment directly by using target detection algorithms, a method is proposed based on improved UNet3+ for green walnut image segmentation, which incorporates the channel and spatial attention mechanism CBAM (convolutional block attention module) and cross-entropy loss function (cross-entropy loss) into the UNet3+ network structure, and introduces the five-layer CBAM in the encoder module to construct the improved UNet3+ network model. The model consists of an encoder module (down-sampling), a decoder module (up-sampling) and a full-scale skip connection module, a full-scale feature supervision module, and a classification guidance module. After utilizing data-enhanced approaches to expand the green walnut dataset, the improved UNet3+ model was trained. The experimental findings demonstrate that the improved UNet3+ network model achieves 91.82% average precision, 96.00% recall rate, and 93.70% F1 score in the green walnut segmentation task; the addition of five-layer CBAM boosts the model segmentation precision rate by 3.11 percentage points. The method can precisely and successfully segment green walnuts, which can serve as a guide and research foundation for precisely identifying and localizing green walnuts and finishing the autonomous sorting for intelligent robots.



**Citation:** Tie, J.; Wu, W.; Zheng, L.; Wu, L.; Chen, T. Improving Walnut Images Segmentation Using Modified UNet3+ Algorithm. *Agriculture* **2024**, *14*, 149. <https://doi.org/10.3390/agriculture14010149>

Academic Editor: Wei Ji

Received: 11 December 2023

Revised: 8 January 2024

Accepted: 15 January 2024

Published: 19 January 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/>).

**Keywords:** image segmentation; green walnut; UNet3+; CBAM

## 1. Introduction

In China, walnut is widely distributed and planted in more than 20 provinces (districts); of these, Xinjiang has the largest walnut planting area at 350,000 hectares, with an annual production of nearly 700,000 tons, ranking second in the nation; Shanxi has the second-largest walnut planting area at 680,000 hectares, with an output of 200,000 tons, ranking third in the nation [1]. However, at present, in many places, like Aksu in Xinjiang, Qingyang in Gansu, Qinling in Shanxi, and Daliang Mountain in Sichuan, the primary method of harvesting walnuts is manual hand-picking and knocking with a long pole which is labor-intensive and inefficient. Walnut quality will be affected if harvesting is not completed in time. Even though there are many different brands of agricultural machinery and numerous technical applications, China's walnut-picking technology is still in its infancy due to its ineffective machinery, its propensity to damage walnut branches, its identification system's shortcomings, which make it difficult to identify the fruit, and other more significant problems. Additional research and development will also be required, which will cost more money [2]. The green walnut fruits growing on the tree in the natural environment will be affected by many interference factors, mainly including the walnut target and the leaves, branches, and other backgrounds between the color being extremely similar, and at the same time due to changes in light and leaf shading, fruit overlap, shadow coverage, and

other factors. Target detection algorithms will have higher leakage or misdetection rates when used to directly identify green walnuts.

In order to recognize walnut targets more accurately, walnut targets must first be distinguished from the background and then detected. Green walnuts have a more complex growing environment, and segmentation can provide richer semantic information to better understand the context and structure of the target and improve the accuracy of detection. At present, there are few studies on the segmentation of green walnuts at home and abroad, and there is an urgent need to propose effective solutions. Therefore, a solution for the segmentation of green walnuts was sought by studying different fruits and vegetables. For example, Zhang Yanfei et al. [3] proposed an apple recognition method based on a two-stage segmentation algorithm that combines region-labeled gradient Hough Circle Transform with Otsu and Watershed. The recognition accuracies of the method were 90.75% in smooth conditions and 89.79% in backlight conditions. Wang Zhifen et al. [4] proposed a new kernel density estimation optimized clustering segmentation algorithm, which obtains the super-pixel region representation of the image by SLIC algorithm, and the local densities of the data points are obtained by kernel density estimation, which drastically reduces the computational amount of the algorithm, and achieves an efficient and accurate segmentation of the target image. Long Jinhui et al. [5] proposed a constrained clustering segmentation algorithm for citrus fruit images based on quantum particle swarm; this algorithm has a lower segmentation error rate as well as a higher peak signal-to-noise ratio than the OTSU algorithm, fuzzy clustering image segmentation algorithm, and other image segmentation algorithms, which reduces the segmentation error rate of the robot image processing system of picking fruit and vegetables. Xu Zhibo et al. [6] proposed a HED optimization network model based on the VGG network and HED network, whose ODS reaches 0.765 on BSDS500, which improves the operational efficiency and adaptability to the complex environment of a pepper-picking robot. Liu Xiaoyang et al. [7] proposed a fruit segmentation method for apple-picking robots based on super-pixel features; this method has better segmentation results with image segmentation accuracy up to 0.9214 and recall up to 0.8565 as compared to the chromatic aberration method using pixel-level features and the fruit segmentation method using neighborhood pixel features. Xu Liming et al. [8] proposed a method based on a homomorphic filtering algorithm for image segmentation of Yangmei, which can effectively segment Yangmei from the background, and the mean value of segmentation error  $A_f$  is only 2.26%, which solves the problem that it is difficult to accurately segment the image of Yangmei fruits in the natural environment. Xu Liming et al. [8] proposed a method based on the homomorphic filtering algorithm for image segmentation of Yangmei, which can effectively segment Yangmei from the background, and the mean value of segmentation error  $A_f$  is only 2.26%, which solves the problem that it is difficult to accurately segment the image of Yangmei fruits in the natural environment. Wang Yude et al. [9] proposed an image segmentation algorithm that fuses color features and texture features to solve the problem of segmenting melon fruit and background images in complex backgrounds. Zhang Hongqi et al. [10] introduced the region information near the threshold point into the segmentation algorithm for the characteristics of tomato fruit images and proposed an improved Otsu threshold segmentation method based on a two-dimensional histogram, which improved the segmentation effect of tomato fruit images in the tomato-picking machine vision system. Xu Liming et al. [11] proposed a segmentation method for Yangmei images based on homomorphic filtering and the K-mean clustering algorithm, and the average values of segmentation error, false-positive rate, and false-negative rate of this algorithm were 3.78%, 0.69%, and 6.8%, respectively. Anindita Septiarini et al. [12] proposed a contour-based automatic image segmentation method for oil palm fruits.

In recent years, deep learning methods have developed rapidly in the field of image segmentation, and scholars at home and abroad have achieved more results in the field of fruit image segmentation. Fan Xiangpeng et al. [13] designed a machine vision detection method based on a deep convolutional neural network in order to realize the recognition and

localization of green walnuts in a walnut garden, which added batch normalization processing and region of interest calibration algorithms on the basis of the Faster R-CNN model, and the improved algorithm modeled a 5.19% increase in accuracy rate compared with the original Faster R-CNN model. Huang Leilei et al. [14] proposed a multi-stage segmentation and morphological recovery method based on deep learning; this method has a fruit recognition AP of 93.66% and a segmentation AP of 96.30%, which solves the problem of occlusion and overlapping of citrus fruits in natural scenes. Liu Changyong et al. [15] researched and proposed a method for automatic segmentation of fruit hearts based on the TMU-Net network by incorporating the Transformer encoder into the U-Net network structure, constructing an improved U-type convolutional network TMU-Net model, and freezing a specific network layer using the migration learning method, which achieves an accuracy rate of 96.72% in the fruit heart segmentation task. Peng Hongxing et al. [16] proposed a full convolutional neural network-based litchi semantic segmentation algorithm, which combines DeepLabV3+ semantic segmentation model and Xception depth-separable convolutional features, with an average intersection and merger ratio (MIOU) of 0.765, which is 0.144 higher than the MIOU of 0.621 of the original DeepLabV3+ model. Jia Weikuan et al. [17] proposed an optimal segmentation algorithm for green fruits based on SOLO, using the separation attention network to design the backbone network of the SOLO algorithm, introducing the feature pyramid network, and constructing the combined structure of ResNeSt+ FPN; the optimal SOLO segmentation algorithm achieves an average recall and precision rate of 94.84% and 96.16%, respectively. Kang Hanwen et al. [18] proposed a multitasking network DaSNet-v2 for visual perception in orchards, realizing the accurate segmentation of apples' IOU of 0.863. The above method studies the recognition of apples, litchi, etc., but there is very little research at home and abroad on green walnuts. For fruit growers who plant a large number of green walnuts and are in urgent need of automated picking solutions, the above proposed method can provide a reference basis for the research on green walnuts.

The UNet family of models is widely used, and for data types not seen in network models, Unet3+ has more image detail information and better generalization ability than the UNet model [19]. UNet3+ is an improved FCN with an encoder–decoder structure, where the feature fusion is no longer conducted by directly adding the pixel values of the corresponding positions of the feature map in the FCN, but by splicing and fusing the feature maps, which increases the number of channels and thus improves the image segmentation precision [20,21]. Compared with UNet and UNet++, UNet3+ redesigns the jump connection using multiscale features and fuses the deep and shallow features of the data through a full-scale jump connection [22–24], which fully extracts the data features with higher recognition accuracy, requires fewer parameters, and produces more accurate location-aware and boundary-enhanced segmentation maps.

Thus, this paper proposes an improved UNet3+ network that introduces the five-layer CBAM in the encoder module and integrates the channel and spatial attention mechanism CBAM (convolutional block attention module) and cross-entropy loss function (cross-entropy loss) into the UNet3+ network structure. This model enables the efficient division of green walnuts in their natural habitat and offers technical assistance for the intelligent harvesting of green walnuts.

## 2. Materials and Methods

### 2.1. Data Collection and Dataset Production

#### 2.1.1. Image Acquisition

This experiment used the green walnut image dataset, partly from the orchard in Zhengning County, Qingyang, Gansu Province, which were obtained via manual shooting by hand with a cell phone, and partly through collecting from major plant websites such as Chinese Field Herbarium, Plant Photo Bank of China, The Royal Horticultural Society, and Chinese Flowers, Plants, and Trees. The collected images have different light intensities, different fruit sizes, or different numbers of fruits, after which the eligible images are subjected to data processing. The dataset contains five categories of image data, including

different degrees of branch and leaf shading, different degrees of fruit overlap, different light intensities, and different fruit sizes and different fruit numbers, etc. The same proportion of images were in each category. Among them, different degree of foliage shading refers to the area of green walnut fruits by foliage shading the fruit area less than 75%; different degree of fruit overlapping refers to the area of green walnut fruits and fruits shading each other less than 75% of the area of the fruits; different intensity of light refers to the different light of the sunlight on the green walnuts; the shooting scene is taken as a sunny day and cloudy day, and the shooting time is taken as 9 o'clock, 14 o'clock, and 17 o'clock; different sizes of the fruits refer to the different sizes of the fruits in the image; and the number of different fruits refers to a variety of the number of fruits in the image, which is taken as (1–6), (6–12), and (more than 12). Figure 1 shows the original green walnut images.



**Figure 1.** Original image of green walnut.

#### 2.1.2. Data Enhancement and Dataset Production

Training deep learning network model parameters provides sufficient sample data. In order to increase the model's capacity for generalization, data enhancement techniques such as image rotation, left-right swapping, zooming in and out, small block distortion, random brightness enhancement/decrease, random color/contrast enhancement/decrease, random clipping, random cropping, random flipping, random grayscale coefficients, random fogging, and Gaussian noise are applied to the original image of the captured green walnut. A total of 10,143 images were obtained through data enhancement. Among them, the probability of image rotation occurring is 80%, the probability of image left-right swap occurring is 50%, the probability of image zoom-in and zoom-out is 80%, and the random luminance enhancement/decrease, random color/contrast enhancement/decrease can be based on the change factor to determine the degree of change. These data enhancement techniques were applied in the order in which they were applied, and Figure 2 displays the sample image obtained—the data-enhanced green walnut image.

In this paper, the image labeling tool Labelme was used to label the green walnut region. An 8-bit grayscale labeled image was generated after processing, where the value of each pixel point represents the species to which the pixel belongs. The expanded sample images were divided in the ratio of 7:2:1 to obtain the training set, validation set, and test set, respectively.



**Figure 2.** Image of green walnut after data enhancement.

## 2.2. Improved UNet3+ Model

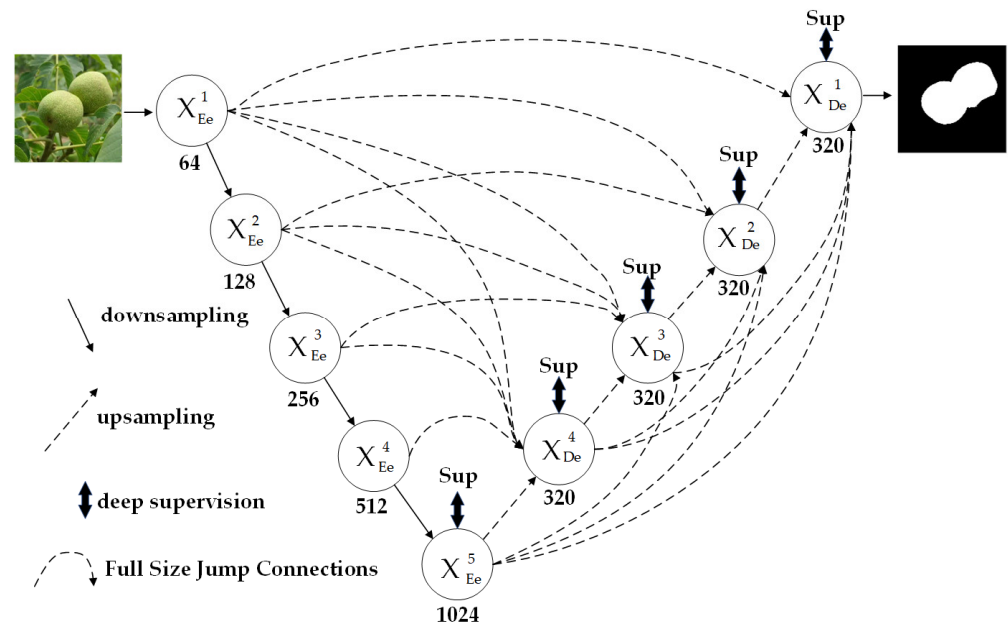
UNet3+ has made several improvements on the basis of the UNet model, but there are still certain issues with it. The UNet3+ model is more complex than the traditional UNet model, requiring more time and computational resources for training; moreover, due to the complexity of the model and the increase in the number of parameters, the UNet3+ model needs more training data to achieve better results; for some small targets or areas where the detail information is not obvious enough, the performance of UNet3+ model still needs to be improved.

### 2.2.1. UNet Network Architecture

The UNet network structure consists of an encoder and a decoder. The encoder consists of multiple convolutional layers and pooling layers for extracting the image features after reducing the size of the feature map. Each convolutional layer consists of convolutional operations, batch normalization, and activation functions to enhance the representativeness of the features and nonlinearities, and the pooling layer is used to reduce the dimensionality of the feature map. The decoder consists of multiple up-sampling layers and convolutional layers for mapping the features into the dimensions of the original image. The up-sampling layer uses an inverse convolution operation to restore the feature map dimensions to the original dimensions. The convolutional layers are used for feature fusion to extract features through a sliding window of the convolutional kernel. In addition, each layer in the decoder section is connected to the corresponding layer in the encoder section to form a jump connection, which passes the low-level features and high-level features in the encoder to the decoder, which can utilize different levels of feature information simultaneously to improve the accuracy of segmentation.

### 2.2.2. UNet3+ Network Architecture

UNet3+ [25–30] consists of five main parts: the encoder module (down-sampling), the decoder module (up-sampling), the full-scale skip connection module, the full-scale feature supervision module, and the classification guidance module. Figure 3 shows the network structure.



**Figure 3.** UNet3+ network structure.

**Encoder module:** The encoding part is the same as that of UNet. First, the input image is convolved twice with  $3 \times 3$  convolution, followed by BatchNorm2d, ReLU. After that, the max pool operation is performed, that is,  $2 \times 2$  convolution with stride = 2. No more down-sampling is performed after the fifth layer of convolution (max pool operation). Whereas,  $3 \times 3$  convolution operation affects the feature channel, down-sampling (Maxpool for down-sampling) affects the resolution. BatchNorm2d performs data normalization. ReLU, activation function. Maxpool uses a  $2 \times 2$  convolution kernel to extract features to shrink the feature map resolution by a factor of 1.

**Decoder module:** The encoder uses pooling and convolution to reduce the size of the image, which reduces the image resolution and causes some detail information to be lost. In the decoding path, the image information is complemented to some extent by twice up-sampling, which restores the image to its original size in order to classify each pixel point. Each decoder layer fuses small-scale feature maps from the encoder, same-scale feature maps, and large-scale feature maps from the decoder, and the feature maps capture fine-grained semantics and coarse-grained semantics at full scale. The small-scale feature maps from the encoder reduce the size of the feature maps by maximum pooling and change the number of channels of the feature maps by convolution. Same-scale feature maps from the encoder change the number of channels of the feature map by convolution. Large-scale features from the decoder expand the size of the feature map by up-sampling and change the size and number of channels of the feature map by convolution. After that, all the feature maps are spliced through the channel level to obtain a feature map with a new number of channels, and then convolution, BN + ReLU, and after that, the feature maps of the decoding layer are obtained to realize the full-scale feature fusion.

**Full-scale jump connectivity module:** Full-size jump connectivity changes the interconnections between the encoder and decoder and the internal connections within the decoder. Each decoder layer in UNet3+ incorporates small-scale and same-scale feature maps from the encoder and large-scale feature maps from the decoder, which capture fine-grained and coarse-grained semantics at the full scale. In this way, it compensates for the lack of UNet

and UNet++ to explore enough information from the full scale to obtain a clear image of the location and boundary of the target. Each decoder layer  $X_{De}^i$  fuses feature maps from different sources as Expression (1) shows.

$$X_{De}^i = \left\{ \begin{array}{l} X_{En}^i, \\ H \left( \left[ \underbrace{C \left( D \left( X_{En}^k \right) \right)_{k=1}^{i-1}}_{Scales: 1^{th} \sim i^{th}}, C \left( X_{En}^i \right), \underbrace{C \left( U \left( X_{De}^k \right) \right)_{k=i+1}^N}_{Scales: (i+1)^{th} \sim N^{th}} \right] \right), \end{array} \right. \quad \left. \begin{array}{l} i = N \\ i = 1, \dots, N-1 \end{array} \right\} \quad (1)$$

$C(\cdot)$  presents a convolution operation,  $H(\cdot)$  represents a feature propagation mechanism including a convolution, a batch normalization, and a ReLU activation function.  $D(\cdot)$  and  $U(\cdot)$  represent down-sampling and up-sampling, and  $[\cdot]$  represent concatenation operations.

**Full-scale feature supervision module:** Full-scale deep supervision is proposed on UNet3+ to generate a broadside output supervised by ground truth at each decoding layer. This step involves the following operations:  $3 \times 3$  conv, bilinear up-sampling, and sigmoid. The specific operation of deep supervision: The last layer of the feature map generated by the feature aggregation mechanism of each decoding layer is fed into the  $3 \times 3$  convolutional layer, which is then accompanied by a bilinear up-sampling, where the up-sampling is, in this case, to restore the resolution of the feature map to the level of the input image. The output of deep supervision is then obtained by multiplying the segmentation result obtained after up-sampling by the classification module's result 0/1. The result of this multiplication is then subjected to sigmoid processing. And the result obtained is the output of deep supervision. The deeply supervised result is then input into the loss function.

**Classification bootstrap module:** To achieve more accurate segmentation results, UNet3+ predicts whether the input image contains the segmented target or not by adding an additional classification task. As Figure 4 shows, the deepest 2D tensor Encoder5 undergoes a series of operations containing Dropout, Convolution, Maxpooling, and Sigmoid, and ends up with two values representing the probability of having/not having a segmentation target. Utilizing the richest semantic information, the classification results can be further guided by outputting two steps for each cut side. With the help of the Argmax function, the two-dimensional tensor is transformed into a single output of {0,1} representing the presence/absence of a segmentation target, with 0 representing absence and 1 representing presence. Subsequently, the individual classification outputs are multiplied with the side segmentation outputs. Due to the simplicity of the binary classification task, this module obtains accurate classification results by optimizing the binary cross-entropy loss function.

### 2.2.3. Improved UNet3+ Network Architecture

In this paper, we propose an improved UNet3+ network that incorporates CBAM (channel and spatial Attention) and cross-entropy loss function (cross-entropy loss). The model introduces CBAM (channel and spatial attention) in the encoder module, focuses on the edge features of the green walnut image, suppresses unnecessary regional responses, and suppresses irrelevant noise information such as illumination, etc. Figure 5 shows the structure of the improved UNet3+ network.

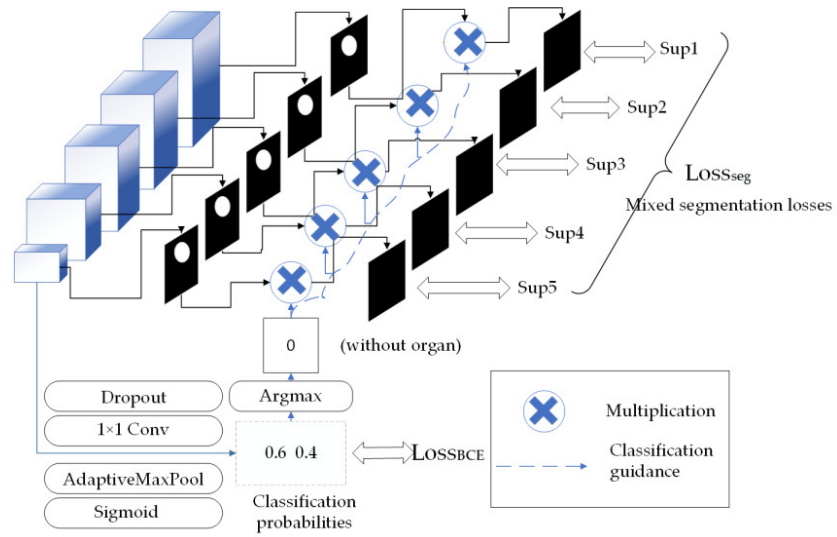


Figure 4. Classification guidance module.

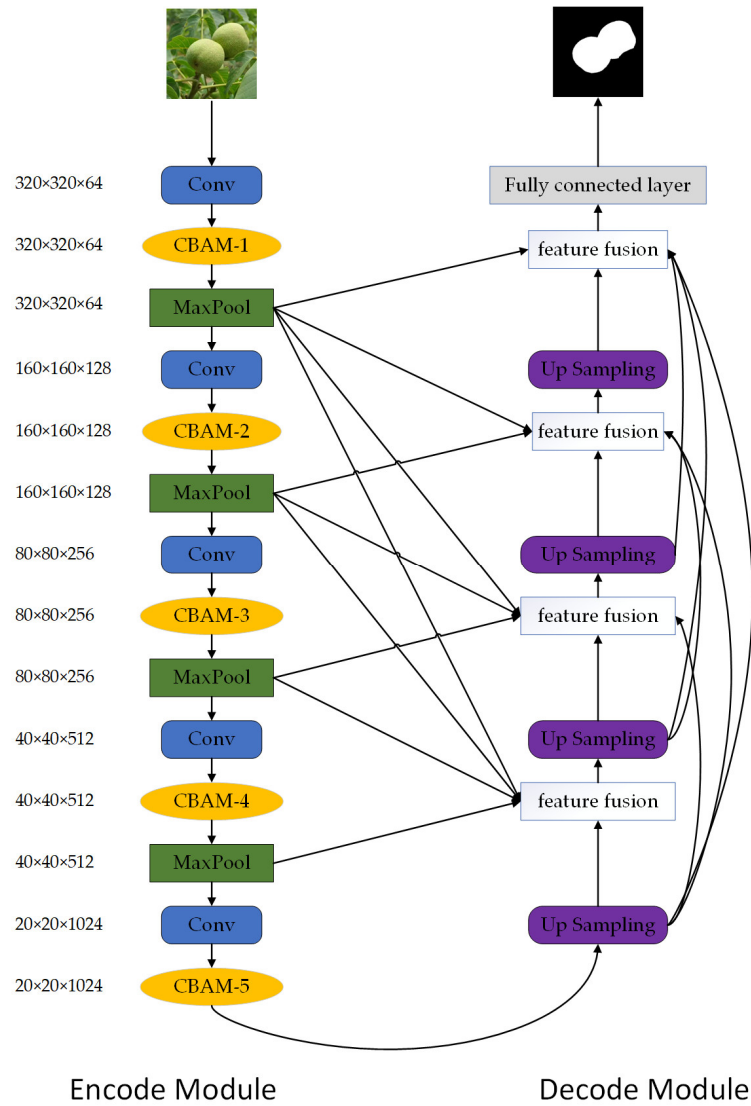


Figure 5. Improved UNet3+ network architecture.



CBAM is introduced in the encoder module to focus on the edge features of the green walnut image and suppress the unnecessary regional response to the uncorrelated light noise information. CBAM is a kind of channel and spatial attention in hybrid attention. As Figure 6 shows for CBAM [31], the CBAM module is able to sequentially generate attention feature map information in both channel and spatial dimensions, and then the two kinds of feature map information are multiplied with the previous original input feature map for adaptive feature correction to produce the final feature map of the green walnut image. The multiplication signs in Figure 6 denote element-level multiplication, with a broadcast mechanism for dimensional transformation and matching in between. The first multiplication sign indicates that the input feature  $F$  is multiplied with the output  $M_c(F)$  of the channel attention module to obtain the result  $F'$ . The second multiplication sign indicates that the result  $F'$  after multiplying the input feature with the output of the channel attention module enters the spatial attention module to obtain  $M_c(F')$ , and  $M_c(F')$  is multiplied with  $F'$  to obtain  $F''$ . The network backbone generates the feature map  $F \in \mathbb{R}^{C \times H \times W}$ , and the CBAM produces the 1D channel attention feature map  $M_c \in \mathbb{R}^{C \times 1 \times 1}$ , and the 2D spatial attention feature map  $M_s \in \mathbb{R}^{1 \times H \times W}$ . The computational formula is:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \quad (2)$$

$\otimes$  denotes element-level multiplication, with a broadcast mechanism for dimensional transformation and matching in between.

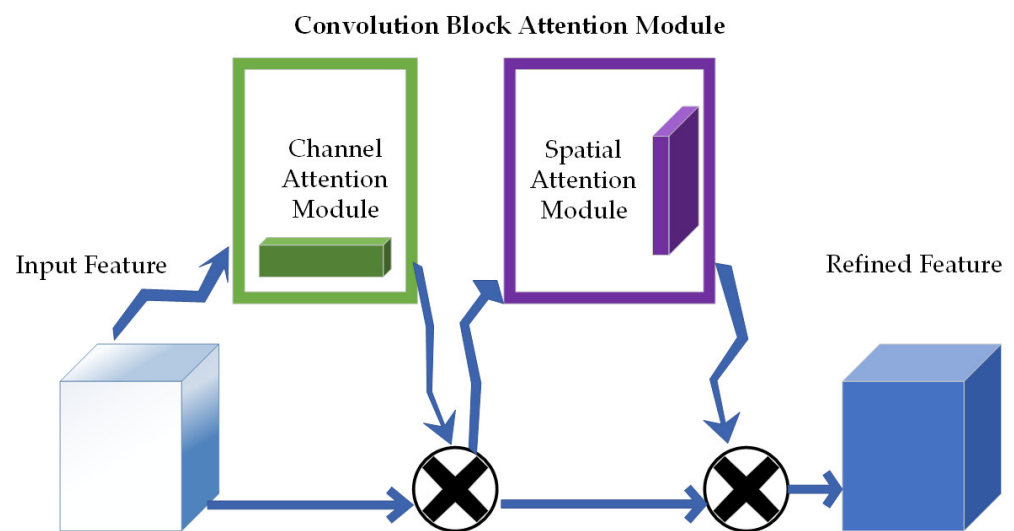


Figure 6. CBAM network architecture.

As Figure 7 shows for the channel attention module, both average and maximum pooling methods are used in the channel attention module;  $F_{avg}^c$  and  $F_{max}^c$  represent the average pooled feature and maximum pooled feature, respectively. The features are then fed into a shared multilayer perceptron (MLP) network to generate the final channel attention feature map  $M_c \in \mathbb{R}^{C/r \times 1 \times 1}$ , and the channel attention is computed as:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W1(W0(F_{avg}^c)) + W1(W0(F_{max}^c))) \end{aligned} \quad (3)$$

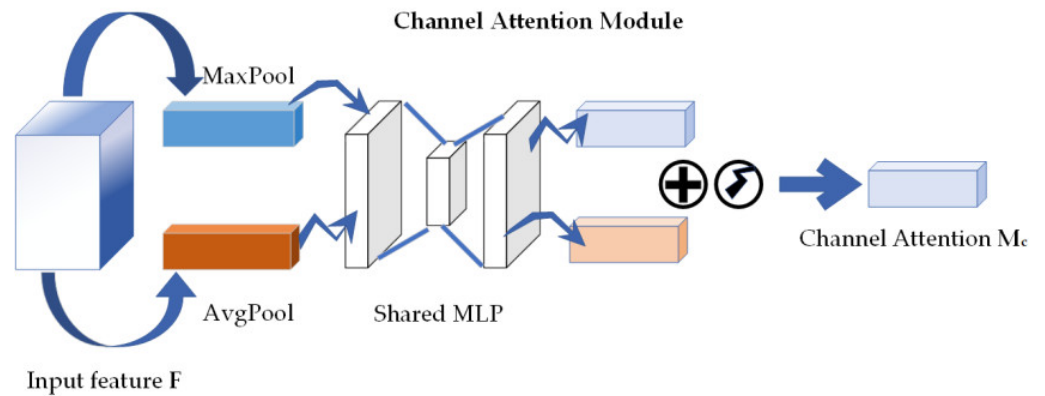


Figure 7. Channel attention module.

Figure 8 shows the spatial attention module. To compute the spatial attention, firstly, average pooling and maximum pooling in the channel dimension are performed, and then the feature maps they produce are spliced together (concat). Then on the spliced feature maps, a convolution operation is used to generate the final spatial attention feature maps  $M_s(F) \in R^{H,W}$ . Two pooling methods are used in the channel dimension to generate the 2D feature maps:  $F_{avg}^s \in R^{1 \times H \times W}$  and  $F_{max}^s \in R^{1 \times H \times W}$ , and the spatial attention computation formula is given as follows:

$$M_s(F) = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) = \sigma(f^{7 \times 7}(F_{avg}^s; F_{max}^s)) \quad (4)$$

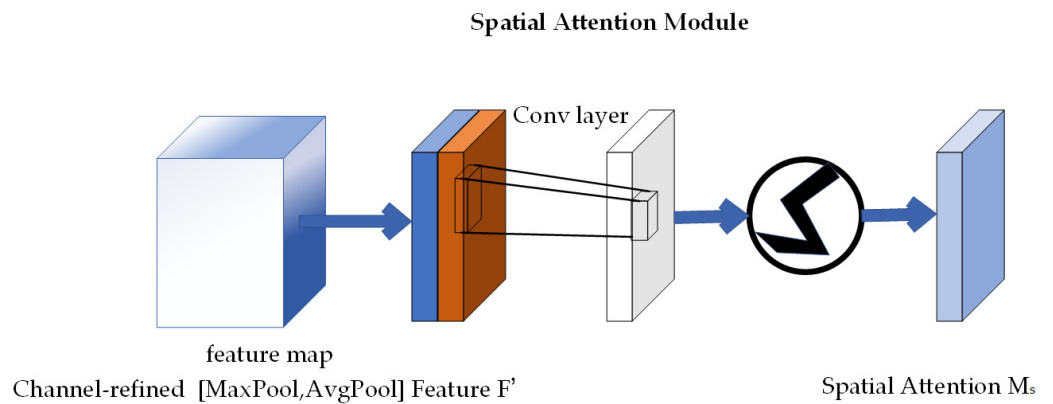


Figure 8. Spatial attention module.

#### 2.2.4. Constructing the Loss Function

The green walnut dataset for this experiment is class-balanced and pixels in the image can be learned equally. Cross-entropy loss [32] is used as the loss function of the model and can improve the performance of network models. The cross-entropy loss function is calculated as:

$$L_{CE} = - \sum_{c=1}^M y_c \lg(p_c) \quad (5)$$

where  $L_{CE}$  is the cross-entropy loss;  $M$  is the number of categories of the sample;  $y_c$  has only two values 0 and 1, if the sample category is consistent with the current category,  $y_c$  takes 1, otherwise  $y_c$  takes 0;  $P_C$  is the probability that the current sample prediction belongs to the  $c$  category. The loss function is a pixel-level cross-entropy loss, which examines each pixel individually and compares the predictions (probability distribution vectors) for each pixel category with the coded label vectors.

### 2.3. Experimental Design

#### 2.3.1. Training Environment

Hardware environment: The algorithm processing platform is a VMware 7.1 server, the processor is an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, the graphics card model is an NVIDIA Corporation GP102GL [Tesla P40], and the GPU model is a VMware SVGA II Adapter.

Software environment: The Ubuntu 20.04.1 LTS was used, the programming language was Python 3.8.17, and the deep learning framework Pytorch 1.7.1 was used for network construction.

#### 2.3.2. Training Parameters and Methods

In this paper, the network optimizer is selected as SGD optimizer, the learning rate is set to  $10^{-3}$ , the number of batch input samples is 2, the loss function is set to CrossEntropyLoss, and the learning rate iterator adopts CosineAnnealingLR. The weight file is saved once for every ten epoch iterations in the experiments, and the training is conducted for a total of 80 rounds. For training the model, the image pixels were  $512 \times 512$  pixels, and the training process took almost 50 h.

#### 2.3.3. Indexes for Model Evaluation

The results in the semantic segmentation task can be categorized as true positive (TP), false positive (FP), true negative (TN), false negative (FN). Where negative refers to the part of the non-object label (generally for the background information other than the target), and positive generally refers to the part of the information that contains the label. TP refers to the part of the input image that contains the label information and is correctly recognized as the corresponding label information; FP refers to the part of the input image that contains the background information and is incorrectly recognized as the label information; TN means the part of the input image containing background information and is correctly recognized as the corresponding background information; FN means the part of the input image containing label information and is incorrectly recognized as background information. Table 1 shows the corresponding confusion matrix.

**Table 1.** Confusion matrix.

		Real Value	Real Value
		Positive	Negative
Predicted value	Positive	True positive	False positive
	Negative	False negative	True negative

Intersection Over Union (IOU), Precision, Recall, and F1-score are used as model evaluation indexes. The Intersection Over Union is the ratio of the overlapping part of the predicted and real regions to the pooled part. The formulas for Precision and Recall are as follows:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (7)$$

The Precision is the ratio of the number of correctly predicted positive samples to the number of samples predicted to be positive, that is, the “accuracy rate”; Recall is the ratio of the number of correctly predicted positive samples to the number of true positive samples, that is, the “check-perfect rate”. The F1-score is a harmonic mean that combines precision and recall, and is calculated as follows:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

### 3. Results and Discussion

#### 3.1. Impact of CBAM Module on Model Performance

Under the same conditions of other parameters, this paper trains the model before and after using the CBAM module and validates it on the test set. Table 2 shows the experimental results. Compared to not using the CBAM module, the IOU, accuracy, and F1 values of the improved UNet3+ (the lowest layer of the encode module, CBAM) network model segmentation are improved by 0.76, 0.8, and 0.47 percentage points, respectively, which indicates that the CBAM module effectively improves the performance of the network model.

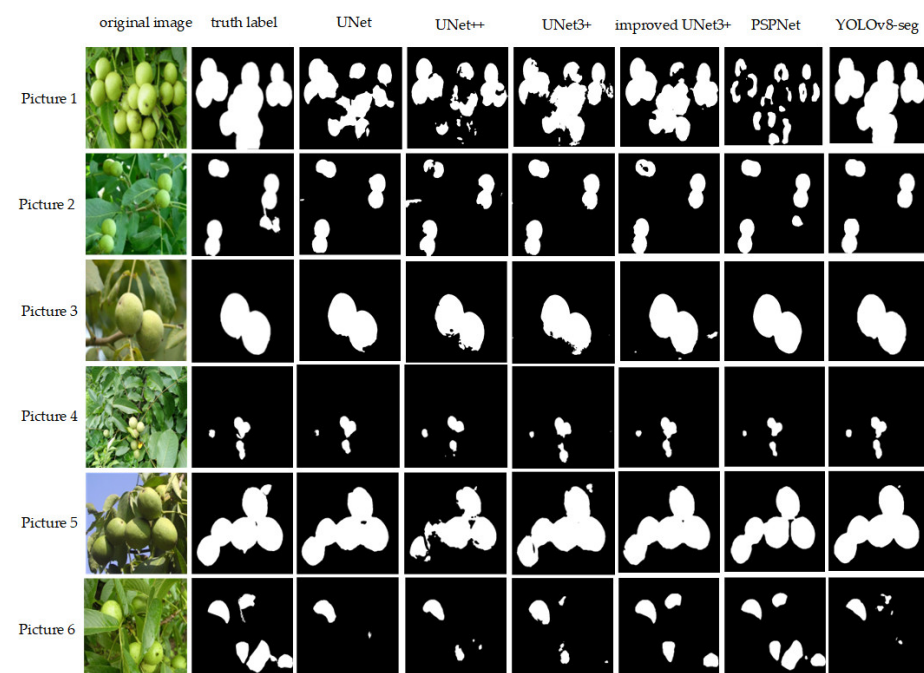
**Table 2.** Experimental results before and after using the CBAM module.

Model	IOU/%	Precision/%	Recall/%	F1-Score/%
Without CBAM	84.89	88.71	95.64	91.60
With CBAM (one layer)	85.65	89.51	95.50	92.07
With CBAM (five layers)	88.36	91.82	96.00	93.70

The IOU, accuracy, and F1 values of the improved UNet3+ (five layers of CBAM are added to the encode module) network model segmentation are improved by 3.47, 3.11, and 2.1 percentage points, respectively, indicating that the increase in the number of CBAM layers effectively improves the performance of the network model.

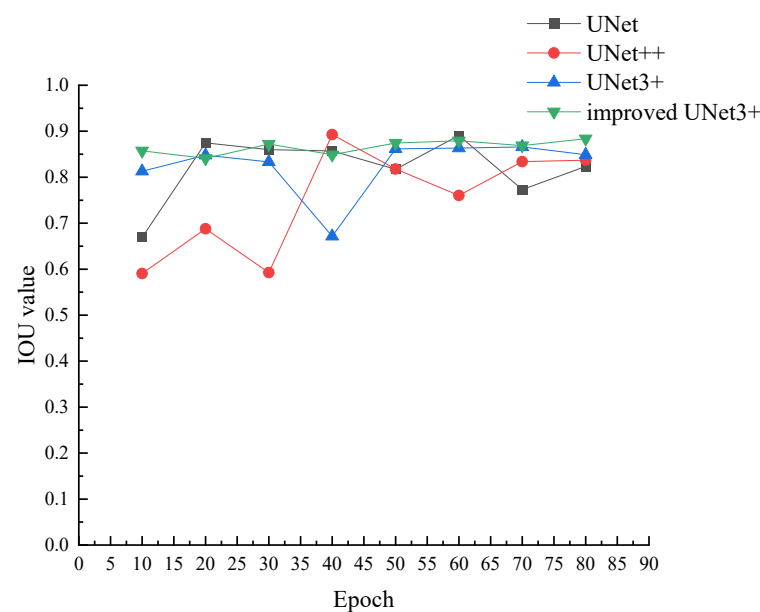
#### 3.2. Improved UNet3+ Network Model Improvement Effect

UNet, UNet++, UNet3+, and PSPNet, as classical models in semantic segmentation, are characterized by high detection accuracy and model innovation. Therefore, in the green walnut segmentation effect experiment, UNet model, UNet++ model, UNet3+ model, PSPNet model, and improved UNet3+ model are selected to compare the segmentation effect. Figure 9 shows the segmentation effect of green walnuts with different network models, and the real labels are the self-manually labeled labels. The comparison shows that for the more overlapping green walnuts, the other models will have obvious segmentation incompleteness, which is difficult to be applied in practical tasks, while the improved UNet3+ network model can segment the green walnuts more accurately and comprehensively, which further shows that the improvements made in this paper are effective.



**Figure 9.** Segmentation effect of green walnut with different network models.

In order to comprehensively analyze the segmentation performance differences between the improved UNet3+ model and the UNet model, UNet++ model, and UNet3+ model, under the same experimental conditions of the other parameters, 80 rounds of training were performed using the green walnut dataset constructed in this paper, and Figure 10 shows the change curves of the training intersection and integration ratios (IOUs) of each model. It is found that, compared with other models, the IOU curve of the improved UNet3+ model's IOU curve is at the top, and the IOU value is finally stabilized at 88.36%, indicating that a better performance is achieved at the early stage of training. As can be seen in Figure 9, in the scenario of overlapping green-skinned walnut fruits, the UNet model and PSPNet model are obviously incompletely segmented, while the improved UNet3+ can segment the boundary better than the other models, as Pictue1 shows for the small-targeted fruits, the UNet3+ model is incompletely segmented and the boundary is not smooth, and the fruit portion of the YOLOV8-seg model is not partially segmented. For small target fruits, the UNet3+ model segmentation is incomplete, and the boundary is not smooth, and the YOLOV8-seg model fruit part segmentation is incomplete, while the improved UNet3+ model segmentation of the fruit contour is very clear, which is better than the other models, as Pictue4 shows. For the case that the shadows obscure the fruits under the light, the UNet++ model segmentation is incomplete, while the improved UNet3+ model still segments out the complete fruit contour, as Pictue5 shows. Therefore, in synthesizing various scenarios, the improved UNet3+ model is clearly superior to the other models.



**Figure 10.** Curve of IOU values with training rounds.

The trained different network models were tested on the green walnut test set images and evaluated based on IOU, Precision, Recall, F1-score, and Time, and Table 3 shows the experimental results.

**Table 3.** Accuracy evaluation of the segmentation effect of different network models on green walnuts.

Network Name	IOU/%	Precision/%	Recall/%	F1-Score/%	Time s/6
UNet	82.33	86.60	95.12	89.98	0.7671
UNet++	83.70	87.58	95.70	90.85	1.4050
UNet3+	84.89	88.71	95.64	91.60	1.3632
YOLOv8-seg	72.08	91.44	77.29	83.77	0.0588
PSPNet	86.35	89.38	96.91	92.49	0.7124
Improved UNet3+	88.36	91.82	96.00	93.70	1.6931

As can be seen from Table 3, the improved UNet3+ model has improved performance in the green walnut segmentation task compared with the UNet model, in which the IOU value, accuracy rate, and F1 value are improved by 6.03, 5.22, and 3.72 percentage points, respectively; compared with the UNet++ network, the IOU value, accuracy rate, and F1 value are improved by 4.66, 4.24, and 2.85 percentage points, respectively; compared with the UNet3+ model, the accuracy rate was improved by 3.11 percentage points; and compared with the PSPNet model, the accuracy rate was improved by 2.44 percentage points. Compared with the YOLOv8-seg model, although the precision is only improved by 0.38 percentage points, the combined evaluation index F1 of precision and recall is 9.93 percentage points higher, and the improved UNet3+ is more advantageous. When studying the inference time of every six images, the average time consumed by the improved UNet3+ model is 1.6931 s. Upgrading the manual recognition method commonly used at the present stage to intelligent recognition, the detection accuracy is more important than the detection time when considering the performance of different intelligent recognition methods, and the detection time of a single green walnut is order of “seconds”, which can meet the detection requirements, so the detection time of the improved UNet3+ model has met the demand.

#### 4. Conclusions

When aiming for the use of target detection algorithms to directly recognize green walnuts in the natural environment, there will be more problems such as omission or misdetection. In order to meet the demand for accurate recognition of green walnuts, this paper proposes an improved UNet3+ network model. The model introduces the channel and spatial mechanism (convolutional block attention module, CBAM) module in the encoder module, focuses on the edge features of the green walnut image, suppresses unnecessary regional responses, and suppresses irrelevant noise information such as illumination. The model realizes the accurate segmentation of green walnut, provides a technical reference for the automatic detection of green walnut, and provides a new solution to the problem of target detection and recognition in agricultural automation, and the main conclusions are as follows:

1. The CBAM module focuses on the edge features of the green walnut image, suppresses the unnecessary regional response, and suppresses the noise information that is irrelevant, such as illumination, etc., and the model is able to realize accurate segmentation, and the segmentation accuracy of the model is as high as 91.82%, which is an improvement of 3.11 percentage points compared with the UNet3+ model that does not use the CBAM.
2. The research method in this paper is feasible in green walnut image segmentation, and the accuracy, IOU value, and F1 value on the test set reach 91.82%, 88.36%, and 93.70%, respectively, with high extraction accuracy. Compared with other classical deep learning models for green walnut, the improved UNet3+ model shows better extraction effect in green walnut segmentation, and the detection time of a single green walnut is in the order of “seconds” to meet the detection requirements, which can meet the practical application level.

**Author Contributions:** Validation, conceptualization, methodology software, formal analysis, writing—review and editing, J.T. and L.Z.; writing—original draft preparation, conceptualization, methodology, J.T. and W.W.; validation, conceptualization, methodology, formal analysis, W.W.; validation, conceptualization, methodology, formal analysis, L.W.; validation, conceptualization, methodology, formal analysis, T.C.; validation, methodology, formal analysis, W.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is funded by the Special Project on Regional Collaborative Innovation in Xin-jiang Uygur Autonomous Region (Science and Technology Aid Program) under grant number 2022E02035, Hubei Provincial Administration of Traditional Chinese Medicine Research Project on Traditional

Chinese Medicine under grant number ZY2023M064, and Wuhan Knowledge Innovation Special Dawn Project under grant number 2023010201020465.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding author. The data are not publicly available.

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

## References

- Li, Y.; Ma, W.Q.; Zhu, Z.J.; Liu, K.; Tian, X. Status quo of walnut industry development in Xinjiang and countermeasure suggestions. *J. Agronomy* **2019**, *9*, 80–86.
- Meng, J.; Fang, X.P.; Shi, X.M.; Zhang, Y.; Liu, J. China's walnut industry development status quo, problems and suggestions. *China Oil Grease* **2023**, *48*, 84–86+103. [\[CrossRef\]](#)
- Zhang, Y.F.; Liu, M.Y.; Gong, J.L.; Lan, Y.B. Apple Recognition Based on Two-Stage Segmentation and Region-Labeled Gradient Hough Circle Transformation. *J. Agric. Eng.* **2022**, *38*, 110–121.
- Wang, Z.F.; Jia, W.K.; Mou, S.H.; Hou, S.J.; Yin, X.; Ji, Z. Green apple segmentation algorithm based on kernel-optimized density clustering. *Spectrosc. Spectr. Anal.* **2021**, *41*, 2980–2988.
- Long, J.H.; Ding, T. Quantum particle swarm based constrained clustering segmentation algorithm for citrus fruit images. *Jiangsu Agric. Sci.* **2018**, *46*, 205–208. [\[CrossRef\]](#)
- Xu, Z.B.; Huang, Y.; Sun, H.B.; Wan, F.X.; Ma, G.J. Research on Pepper Fruit Segmentation Algorithm in Natural Scene. *For. Mach. Woodwork. Equip.* **2022**, *50*, 73–77. [\[CrossRef\]](#)
- Liu, X.Y.; Zhao, D.A.; Jia, W.K.; Ruan, C.Z.; Ji, W. Fruit segmentation method for apple picking robots based on superpixel features. *J. Agric. Mach.* **2019**, *50*, 15–23.
- Xu, L.M.; Lv, J.D. Research on segmentation methods for waxberry fruit images in natural environment. *J. Shenyang Agric. Univ.* **2016**, *47*, 334–341.
- Wang, Y.D.; Zhang, X.Z. Algorithm for Melon Fruit Segmentation in Complex Background. *J. Agric. Eng.* **2014**, *30*, 176–181.
- Zhang, H.Q.; Liu, Y.; Hao, M. Research on tomato fruit image segmentation method based on machine vision. *Res. Agric. Mech.* **2015**, *37*, 58–61. [\[CrossRef\]](#)
- Xu, L.M.; Lv, J.D. Image Segmentation of waxberry Based on Homomorphic Filtering and K-mean Clustering Algorithm. *J. Agric. Eng.* **2015**, *31*, 202–208.
- Septiarini, A.; Hamdani, H.; Hatta, H.R.; Anwar, K. Automatic image segmentation of oil palm fruits by applying the contour-based approach. *Sci. Hortic.* **2020**, *261*, 108939. [\[CrossRef\]](#)
- Fan, X.P.; Xu, Y.; Zhou, J.P.; Liu, X.D.; Tang, J.S. Walnut recognition and localization based on improved Faster R-CNN. *J. Yanshan Univ.* **2021**, *45*, 544–551.
- Huang, L.L.; Miao, Y.B. Deep learning based overlapping citrus segmentation and morphological recovery. *Res. Agric. Mech.* **2023**, *45*, 70–75. [\[CrossRef\]](#)
- Liu, C.Y.; Li, S.J.; Shi, H.; Zha, Z.H.; Deng, H.T. Apple fruit center segmentation method based on TMU-Net network. *J. Agric. Eng.* **2022**, *38*, 304–312.
- Peng, H.X.; Xue, C.; Shao, Y.Y.; Chen, K.Y.; Xiong, J.T.; Xie, Z.H.; Zhang, L.H. Semantic Segmentation of Litchi Branches Using DeepLabV3+Model. *IEEE Access* **2020**, *8*, 164546–164555. [\[CrossRef\]](#)
- Jia, W.K.; Li, Q.W.; Zhang, Z.H.; Liu, G.L.; Hou, S.J.; Ze, J.; Zheng, Y.J. Optimized SOLO segmentation algorithm for green fruits of persimmons and apples in complex environments. *J. Agric. Eng.* **2021**, *37*, 121–127.
- Kang, H.W.; Chen, C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. *Comput. Electron. Agric.* **2020**, *171*, 105302. [\[CrossRef\]](#)
- Zhong, Z.; Peng, J.Y.; Yu, L.; Shan, M.G. Design and Implementation of Scattering Imaging Experiment Based on UNet3+. *Lab. Res. Discov.* **2023**, *42*, 25–29. [\[CrossRef\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015.
- Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*; Springer: Berlin/Heidelberg, Germany, 2018.
- Zhou, Y.; Zhang, Y.; Yang, D.; Lu, J.; Li, G. Pipeline signal feature extraction with improved VMD and multi-feature fusion. *Syst. Sci. Control. Eng. Open Access J.* **2020**, *8*, 318–327. [\[CrossRef\]](#)
- Vimina, E.R.; Jacob, K.P. Feature fusion method using BoVW framework for enhancing image retrieval. *Image Process. IET* **2019**, *13*, 1979–1985. [\[CrossRef\]](#)
- Hou, S.; Sun, Q. An orthogonal regularized CCA learning algorithm for feature fusion. *J. Vis. Commun. Image Represent.* **2014**, *25*, 785–792. [\[CrossRef\]](#)
- Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.

26. Alam, T.; Shia, W.-C.; Hsu, F.-R.; Hassan, T. Improving Breast Cancer Detection and Diagnosis through Semantic Segmentation Using the Unet3+ Deep Learning Framework. *Biomedicines* **2023**, *11*, 1536. [[CrossRef](#)] [[PubMed](#)]
27. Xu, Y.; Hou, S.; Wang, X.; Li, D.; Lu, L. A Medical Image Segmentation Method Based on Improved UNet 3+ Network. *Diagnostics* **2023**, *13*, 576. [[CrossRef](#)]
28. Liu, Z.; He, X.; Lu, Y. Combining UNet 3+ and transformer for left ventricle segmentation via signed distance and focal loss. *Appl. Sci.* **2022**, *12*, 9208. [[CrossRef](#)]
29. Yin, M.; Wang, P.; Ni, C.; Hao, W. Cloud and snow detection of remote sensing images based on improved Unet3+. *Sci. Rep.* **2022**, *12*, 14415. [[CrossRef](#)]
30. Hou, Y. Research on segmentation of MRI brain tumor image based on improved UNet3+. In Proceedings of the International Conference on High Performance Computing and Communication (HPCCE 2021), Xiamen, China, 3–5 December 2021; pp. 242–247.
31. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Pang, T.; Xu, K.; Dong, Y.; Du, C.; Chen, N.; Zhu, J. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv* **2019**, arXiv:1905.10626.

**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.