


## Article

# Fruit Stalk Recognition and Picking Point Localization of New Plums Based on Improved DeepLabv3+

Xiaokang Chen <sup>1,2</sup>, Genggeng Dong <sup>1,2</sup>, Xiangpeng Fan <sup>3,4,\*</sup> , Yan Xu <sup>1,2,\*</sup>, Tongshe Liu <sup>1,2</sup>, Jianping Zhou <sup>1,2</sup> and Hong Jiang <sup>1,2</sup>

<sup>1</sup> College of Intelligent Manufacturing and Modern Industry, Xinjiang University, Urumqi 830017, China; chenxk@xju.edu.cn (X.C.)

<sup>2</sup> Agriculture and Animal Husbandry Robot and Intelligent Equipment Engineering Research Center of Xinjiang Uygur Autonomous Region, Urumqi 830049, China

<sup>3</sup> Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

<sup>4</sup> Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China

\* Correspondence: fanxiangpeng@caas.cn (X.F.); xuyan2018@xju.edu.cn (Y.X.)

**Abstract:** Among the challenges posed by real orchard environments, where the slender new plum fruit stalks exhibit varying postures and share similar coloration with surrounding leaves and branches, the significant obscuration caused by leaves leads to inaccurate segmentation of the fruit stalks, thereby complicating the precise localization of picking points and other related issues. This paper proposes a method for new plum fruit stalk recognition and picking point localization based on the improved DeepLabv3+ model. Firstly, it employs the lightweight MobileNetv2 as the backbone feature extraction network. Secondly, the Convolutional Block Attention Module (CBAM) is integrated into the decoder to enhance the model's ability to extract key features of the fruit stalks. Moreover, dense atrous spatial pyramid pooling (DenseASPP) is utilized to replace the original ASPP module, thereby reducing segmentation leakage. Finally, a picking point localization method is designed based on a refinement algorithm and an endpoint detection algorithm to meet the specific picking demands of new plum, identifying the endpoints along the skeletal line of the fruit stalks as the optimal picking points. The experimental results demonstrate that the mean intersection over union (MIoU) and mean pixel accuracy (MPA) of the enhanced DeepLabv3+ model are 86.13% and 92.92%, respectively, with a model size of only 59.6 MB. In comparison to PSPNet, U-Net, and the original DeepLabv3+ model, the MIoU improves by 13.78, 0.34, and 1.31 percentage points, while the MPA shows enhancements of 15.35, 1.72, and 1.38 percentage points, respectively. Notably, with the endpoint of the fruit stalk's skeletal structure designated as the picking point for new plums, the localization success rate reaches 88.8%, thereby meeting the requirements for precise segmentation and picking point localization in actual orchard environments. Furthermore, this advancement offers substantial technical support for the research and development of new plum harvesting robots.

**Keywords:** deep learning; semantic segmentation; attention mechanism; picking point location



**Citation:** Chen, X.; Dong, G.; Fan, X.; Xu, Y.; Liu, T.; Zhou, J.; Jiang, H. Fruit Stalk Recognition and Picking Point Localization of New Plums Based on Improved DeepLabv3+. *Agriculture* **2024**, *14*, 2120. <https://doi.org/10.3390/agriculture14122120>

Academic Editor: Roberto Alves Braga Júnior

Received: 23 October 2024

Revised: 19 November 2024

Accepted: 21 November 2024

Published: 22 November 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

New plums, a European variety renowned for its rich nutritional and medicinal properties, are extensively cultivated in the Kashgar and Yili regions of Xinjiang. The maturity period of new plums is relatively short, necessitating rapid completion of the harvesting process; otherwise, significant economic losses may ensue. Currently, the harvesting of new plums primarily relies on manual picking methods [1]. However, with the accelerated aging of the population in China and a growing shortage of labor resources in rural areas, manual harvesting significantly elevates labor costs. Consequently, the research and development of highly automated and intelligent picking robots for new plums has emerged as a pivotal trend in the future of plum harvesting [2,3].

In the unstructured operating environment of orchards, the ability to quickly and accurately identify new plum fruit stalks and determine the optimal picking point is a critical technology for enabling rapid harvesting by new plum-picking robots. The intricate conditions of new plum orchards, characterized by the slender nature of new plum stalks and significant variations in their postures, exacerbate the challenges associated with accurately identifying new plums stalks and locating the optimal picking point [4].

Currently, the methods for identifying and localizing fruit peduncles can be broadly classified into two major types: traditional identification and localization methods, and those based on deep learning techniques [5,6]. Traditional recognition and localization techniques predominantly depend on attributes such as color, shape, and texture to differentiate fruit stalks from their surrounding background. For instance, Bac et al. [7] employed guide lines as visual cues to identify and locate bell pepper peduncles. Xiong et al. [8] implemented a corner point detection algorithm, analyzing the rate of change of corner points in both horizontal and vertical directions to determine the picking point for lychee peduncles. Additionally, Ji et al. [9] utilized 2R-G-B color features for identifying tomato fruits and auxiliary markers, with the intersection of the fitted fruit peduncle curve and auxiliary marker edges serving as the picking point, achieving a success rate of 88.6%. Luo et al. [10] segmented grape bunches based on color features, pinpointing the picking location by examining the relationship between the fruit stalk and the fruit, attaining a location accuracy of 88.3%. In recent years, the rapid advancement of machine vision technology has led to the widespread adoption of deep learning-based methods for fruit recognition and localization. For example, Yu et al. [11] utilized the R-YOLO model to assess the growth posture of strawberries and determine the picking point based on the rotational angle of the identified bounding box axis. Peng et al. [12] enhanced the segmentation accuracy of lychee fruit stalks in orchard environments by incorporating transfer learning and spatial pyramid pooling into the DeepLabv3+ model. Furthermore, Ning et al. [13] employed an improved Mask R-CNN for the recognition and segmentation of grape fruit stalks, using the centers of the horizontal edges closest to the mass center of the fruit stalk image as the final picking points. Additionally, Rong et al. [14] proposed an improved Swin Transformer V2 semantic segmentation model along with a picking point recognition algorithm, achieving a remarkable recognition rate of 97.4%. Yan et al. [15] proposed a lightweight convolutional neural network approach for tea segmentation and picking point localization, incorporating the optimized MobileNetv2 architecture along with the densely connected atrous spatial pyramid pooling (DASPP) module within the MC-DM framework to enhance the accuracy of picking point identification for tea buds. Wu et al. [16] achieved enhanced segmentation accuracy for potato root systems by integrating an improved backbone network with the DeepLabv3+ model, as well as incorporating the CARAFE up-sampling module and the CBAM attention mechanism, with the potential to reach an accuracy of 94.05%. The preceding research has introduced a variety of methodologies aimed at the identification and localization of fruit stems, achieving substantial breakthroughs in the process. In comparison to the conventional recognition and positioning methods utilizing deep learning, these traditional approaches impose stricter requirements on the orchard environment, exhibit poor adaptability to varying conditions, and demonstrate inadequate recognition performance for fruit stems characterized by variable shapes and unstable positions. These limitations render them challenging to implement effectively in practical applications. Currently, in light of the challenges associated with the recognition and positioning of new plum stems addressed in this paper, we have selected a deep learning-based method for the recognition and localization of fruit stems.

New plum fruit stalks exhibit slender forms and variable postures, with their coloration closely resembling that of the surrounding leaves and branches. Consequently, accurately segmenting the fruit stalks within the intricate orchard environment remains a formidable challenge. To address this challenge, this study proposes an enhanced methodology for the identification of new plum fruit stalks and the localization of picking points, utilizing the DeepLabv3+ model. This involves the replacement of the backbone network

and the introduction of dense void space pyramid pooling and a lightweight attention mechanism module, which collectively enhance the model's segmentation accuracy while simultaneously reducing its size. Subsequently, the fruit stalk of a new plum is segmented using the enhanced DeepLabv3+ model, resulting in a segmented image of the fruit stalk. Following this, the image binarization algorithm, skeleton refinement algorithm, and expansion operation are employed to extract the skeleton line from the segmented image. Finally, the endpoint detection algorithm is utilized to identify the endpoints of the skeleton line, designating them as the picking points. This approach ensures the integrity of the fruit stalk during harvesting and maximizes the economic benefits associated with new plums. This study offers critical data support for the rapid harvesting of new plum fruit, demonstrating significant practical application value in the actual context of orchard picking.

The primary contribution of this research is twofold: Firstly, it addresses the challenge of accurately segmenting new plum fruit stalks within the intricate environment of orchards, a task complicated by the slender morphology of the stalks, the diverse range of gestural variations, and the similarity in coloration between the fruit stalks and the trees' foliage and branches. Secondly, this study proposes an innovative localization methodology for plum picking points that is predicated on skeleton refinement and endpoint detection algorithms, effectively addressing the challenge of accurately identifying the plum picking points. Finally, the methodology delineated herein offers a substantial technical foundation for the ongoing advancement of new-plum-picking robotic systems.

## 2. Dataset Construction and Labeling

To authentically replicate the actual working conditions encountered by new-plum-picking robots in orchard environments, the dataset utilized in this study was collected from new plum plantations located in Chabchal County, Ili Kazakh Autonomous Prefecture, China. Data collection occurred between August and September 2023, during the hours of 10:00 to 18:00, utilizing an iPhone 13 as the collection device. To acquire representative image data of fruit stalks within the complex orchard environment, a total of 413 clear images were captured under varying lighting conditions and of diverse postures. To enhance the learning capacity and robustness of the model, data augmentation techniques, including mirroring, rotating, and panning, were applied to the original images. Consequently, the enhanced dataset comprised 1220 images, with an example illustrated in Figure 1. By means of extending and transforming the existing data, novel data samples were generated, thereby augmenting the size and diversity of the dataset. This methodology facilitates the model's capacity to generalize to previously unseen data with greater efficacy. Furthermore, with the aid of data enhancement techniques, it is possible to train more robust models on constrained datasets, thereby enhancing both the accuracy and generalization capabilities of the model. Subsequently, the LabelMe annotation tool was employed to annotate the dataset, focusing exclusively on new plum fruit stalks, with an example of the annotations presented in Figure 2. Finally, the dataset was partitioned into training and testing subsets in an 8:2 ratio.

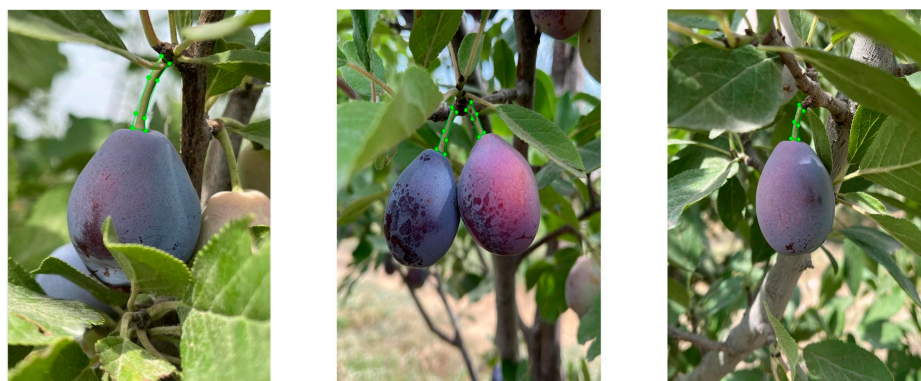


Figure 1. Sample of the new plum dataset.

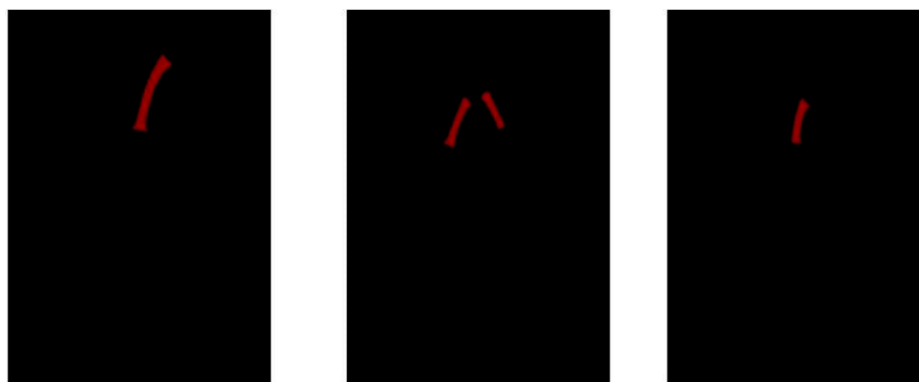


Figure 2. Sample image labeling.

### 3. The Construction of the New Plum Fruit Stalk Segmentation Model and Picking Point Localization

#### 3.1. DeepLabv3+ Modeling

The DeepLabv3+ model [17,18] represents a seminal algorithm within the domain of semantic segmentation. This model utilizes an atrous convolution technique to expand its receptive field, thereby facilitating the acquisition of a broader spectrum of contextual information without incurring additional computational costs. The fundamental architecture of DeepLabv3+ is composed of two primary components: the encoder and the decoder. The encoder component employs the Xception network [19] as its backbone for feature extraction, systematically separating the extracted features into shallow and deep categories. Subsequently, the deep features are supplied to the atrous spatial pyramid pooling (ASPP) module, where the effective extraction of features is achieved through multi-scale convolutions that incorporate varying dilation rates. This procedure empowers the model to concentrate on the high-level features of images while simultaneously addressing the low-level details, thereby enhancing the model's sensitivity to the intricacies and boundaries of the segmented images. The decoder component is tasked with upsampling the deep features from the ASPP output, utilizing fourfold linear interpolation and executing feature fusion with the shallow features extracted from the backbone network. Additionally, another round of quadruple linear interpolation is conducted to reconstruct the feature map, ultimately yielding the segmentation result. The decoder is meticulously designed to facilitate the fine-grained segmentation of images through the cascading of upsampling procedures with the encoder features, thereby allowing the model to pinpoint image boundaries with greater precision, consequently enhancing the overall accuracy of semantic segmentation.

#### 3.2. Improvements to the DeepLabv3+ Model

The DeepLabv3+ model demonstrates effective segmentation of new plum fruit stalks; however, it presents certain challenges. Firstly, the backbone extraction network comprises numerous layers and parameters, leading to increased overall model complexity, which hampers its deployment on mobile embedded devices. Secondly, the model displays reduced performance in accurately segmenting detailed information within images. This paper addresses the aforementioned challenges, as well as the specific issues encountered during the segmentation of new plum fruit stalks, by implementing targeted improvements to the DeepLabv3+ model. Firstly, to address the original DeepLabv3+ model's excessive layers and high number of parameters in the backbone extraction network, this study replaces the original backbone with the improved lightweight MobileNetv2 network [20]. This substitution aims to reduce the parameter count and model complexity, facilitating deployment on mobile embedded devices. Secondly, to mitigate inaccurate segmentation caused by occlusion from branches and leaves, a lightweight attention mechanism, CBAM [21], is incorporated into the decoder section. This enhancement aims to bolster the model's capability to extract features pertinent to the new plum fruit stalks. Finally,

to enhance the model's sensitivity to the edge details of the fruit stalks, the dense atrous spatial pyramid pooling (DenseASPP) [22] module is employed, utilizing dilation rates of 3, 3, 6, 9, and 24. This modification replaces the ASPP module to minimize issues related to leaky segmentation within the model. The structure of the improved model is illustrated in Figure 3.

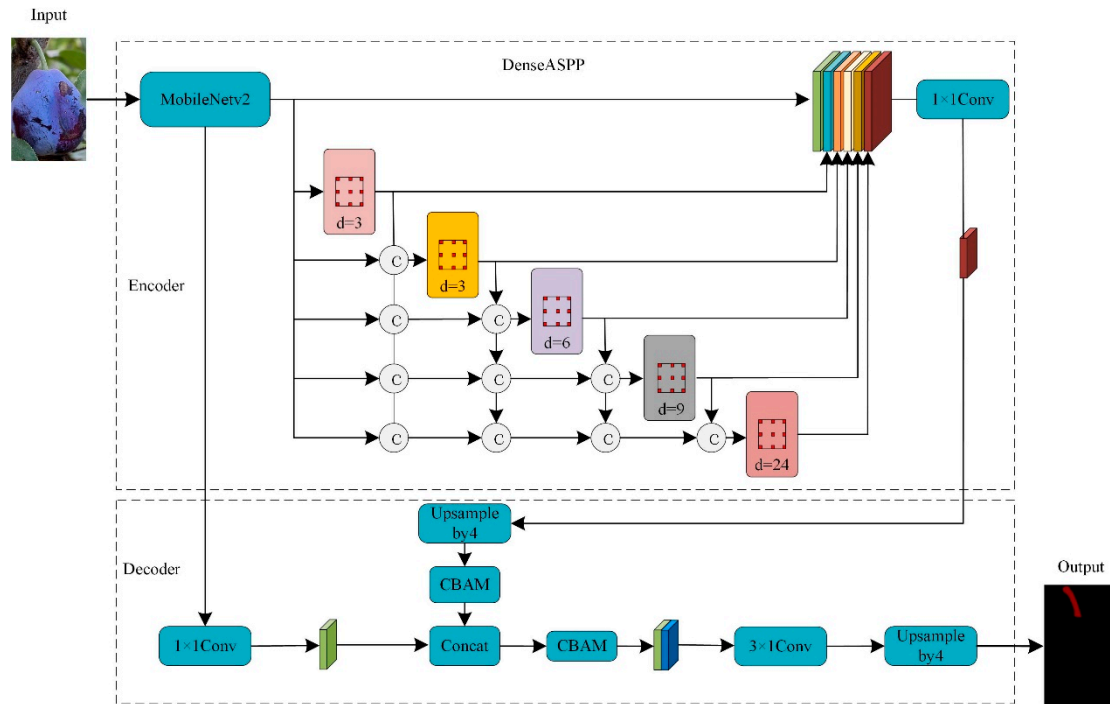
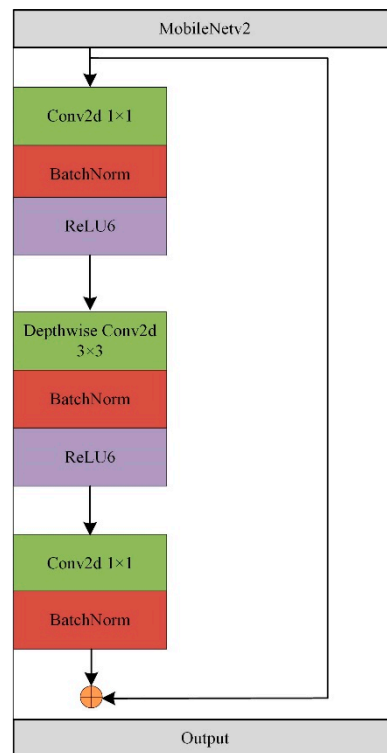


Figure 3. Improved DeepLabv3+ model.

### 3.2.1. Lightweight Feature Extraction Module

As a feature extraction backbone for the DeepLabv3+ model, Xception is characterized by a considerable number of parameters, prolonged training times, and high complexity, which collectively impede its deployment on mobile embedded devices. To mitigate the computational burden associated with the number of parameters and to decrease the model's complexity, this study uses the lightweight MobileNetv2 architecture to supplant the original backbone network. MobileNetv2, introduced by Google's research team in 2018, serves as an effective backbone network module. This architecture is distinguished by its compact size, reduced parameter count, and accelerated computational speed. This framework is specifically designed as a lightweight solution for image segmentation and object detection tasks. The network employs depthwise separable convolution, effectively decoupling spatial and channel convolutions, thereby reducing both the parameter count and the computational complexity. The structural architecture of the network is illustrated in Figure 4. Building upon the foundations of MobileNetv1, MobileNetv2 incorporates a reciprocal residual structure and linear bottleneck. This architecture utilizes  $1 \times 1$  convolutions to increase the input dimensions, followed by  $3 \times 3$  depthwise separable convolutions for feature extraction. Ultimately, a  $1 \times 1$  convolution is employed to reduce the dimensionality of the output. To enhance the integration of the MobileNetv2 architecture within the DeepLabv3+ model, this study proposes further modifications to its network structure. Initially, the architecture employs fourfold downsampling to replace the original fivefold downsampling implemented in the previous network structure, allowing for improved retention of image details. This modification enhances the segmentation quality at the image boundaries and subsequently eliminates the global average pooling and feature classification modules to further minimize the model's parameter count.



**Figure 4.** MobileNetV2 network architecture diagram.

### 3.2.2. CBAM Attention Mechanism

To enhance the model's capability to extract features of new plum stems within the complex orchard environment, this study incorporates the lightweight CBAM attention mechanism in the model's decoder section. The lightweight architecture of the CBAM attention mechanism avoids reliance on extensive convolutional structures, comprising instead a limited number of pooling layers and feature fusion operations. This design effectively mitigates the computational burden associated with convolutional operations, thereby reducing the module's complexity and overall computational demands. The CBAM attention mechanism comprises a channel attention module and a spatial attention module. The architecture of this mechanism is illustrated in Figure 5. Initially, the input feature map undergoes global maximum and average pooling, after which it is processed by both fully connected and convolutional layers, culminating in weight normalization via the sigmoid function. Consequently, both the weighted channel feature map and the spatially weighted feature map are derived. Ultimately, the output features from both the channel attention and spatial attention modules are multiplied elementwise to generate the final attention-enhanced features. The primary function of the channel attention mechanism is to amplify the feature information across each channel, whereas the spatial attention mechanism serves to weigh the importance of varying feature information. To enhance the model's focus on critical feature information, the CBAM attention mechanism effectively augments the model's capacity to represent significant feature details through the integration of both channel attention and spatial attention, thereby improving its performance in image segmentation tasks.

### 3.2.3. Pyramid Pooling Module for Dense Void Spaces

The ASPP module in the DeepLabv3+ model effectively integrates dilated convolution features with varying expansion rates, thereby broadening the network's receptive field without compromising its resolution. This module consists of  $1 \times 1$  convolutions,  $3 \times 3$  convolutions with distinct expansion coefficients, and a global average pooling layer. Through concatenation followed by  $1 \times 1$  convolution, information from multiple branches

is amalgamated to enhance the model’s capability to comprehend multi-scale contexts, increasing the sensitivity field of features without compromising feature resolution, enabling the acquisition of multi-scale image information. However, the ASPP module requires a significant expansion rate to achieve an adequate receptive field when processing new high-resolution images; excessively high expansion rates can lead to dilation convolution failure, resulting in suboptimal image segmentation performance. To address these issues, this study introduces the DenseASPP module, the architecture of which is depicted in Figure 6. The module amalgamates the outputs of individual dilated convolutions through a dense connectivity approach, thereby creating a denser feature pyramid. A sequence of dilated convolutions with varying expansion rates is combined and cascaded to achieve a broader receptive field. This approach effectively circumvents the failure of dilated convolutions that can arise from excessive expansion rates. The target object, the new plum stem, is relatively small; thus, an excessively large expansion rate can result in the loss of critical image details, leading to unsatisfactory segmentation outcomes, particularly at the boundaries of the new plum stem. Consequently, this study modified the three dilated convolution layers in the DenseASPP module, originally set to expansion rates of 6, 12, and 18, to new expansion rates of 3, 6, and 9, thereby enhancing the model’s focus on the details of the new plum stem. The calculation formula for a single receptive field is presented as follows:

$$R_{k,r} = (r - 1) \times (k - 1) + k \tag{1}$$

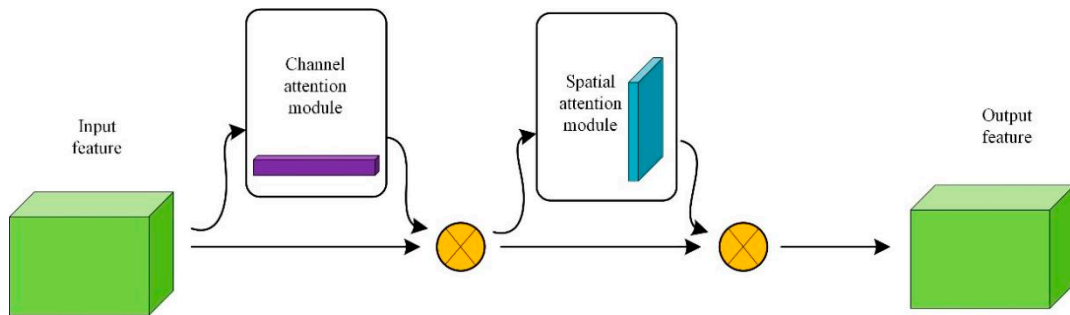


Figure 5. CBAM attention mechanism.

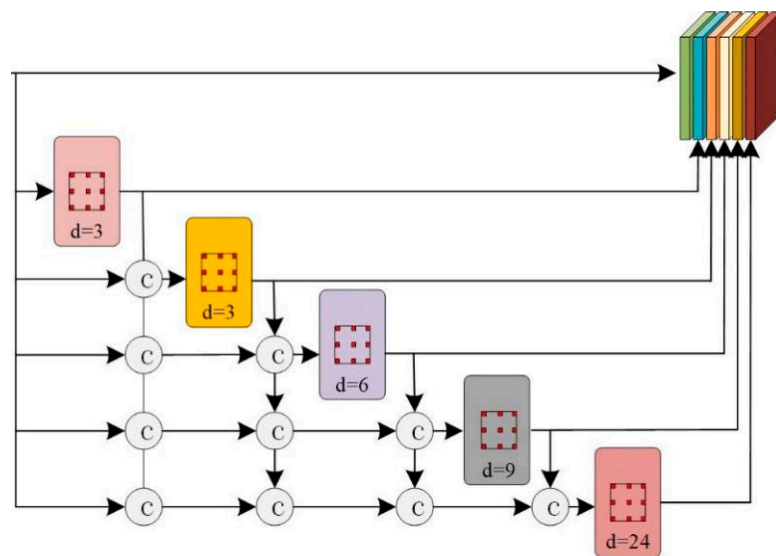


Figure 6. DenseASPP module.

The size of the receptive field for a cascade of two atrous convolutional layers is calculated as follows:

$$R_{k,r} = \sum_{n=1}^N k_n - (n - 1) \quad (2)$$

where  $R_{k,r}$  denotes the size of the receptive field,  $r$  denotes the dilation rate,  $k$  denotes the convolution kernel size, and  $n$  is the number of null convolution layers. ASPP is the parallelization of null convolution layers with different dilation rates, where the maximum receptive field corresponds to the highest dilation rate, with dilation rates of 3, 6, and 9, indicating the extent of the maximum receptive field:

$$R_{max} = \max[R_{3,9}] = 19 \quad (3)$$

DenseASPP obtains a larger receptive field by stacking and connecting the cavity convolutional layers, and the maximum receptive field for expansion rates of (3, 6, 9) is as follows:

$$R_{max} = R_{3,3} + R_{3,6} + R_{3,9} - 2 = 37 \quad (4)$$

### 3.3. Picking Point Location

The localization of the picking point represents a critical step in the operational process of new-plum-picking robots. According to local standards of Xinjiang Uygur Autonomous Region [23], new plums with intact fruit peduncles are classified as high-quality fruits, yielding maximum economic benefits. This study employs fruit stalk refinement and endpoint detection algorithms to accurately identify the picking point. Initially, the improved DeepLabv3+ model is utilized for semantic segmentation of the fruit stalks, effectively isolating new plum stalks from the background before binarizing the segmented image; subsequently, the skeletal line of the fruit stalk is extracted using the refinement algorithm, followed by an expansion operation to compensate for any missing sections of the skeletal line; finally, the endpoint detection algorithm is employed to extract the fruit stalk, and the coordinates of the skeletal line's endpoint are then extracted via the endpoint detection algorithm, designating this endpoint as the final picking point. This approach ensures the maximum integrity of the fruit stalks, thereby optimizing the economic benefits derived from new plums. The operational process is illustrated in Figure 7.

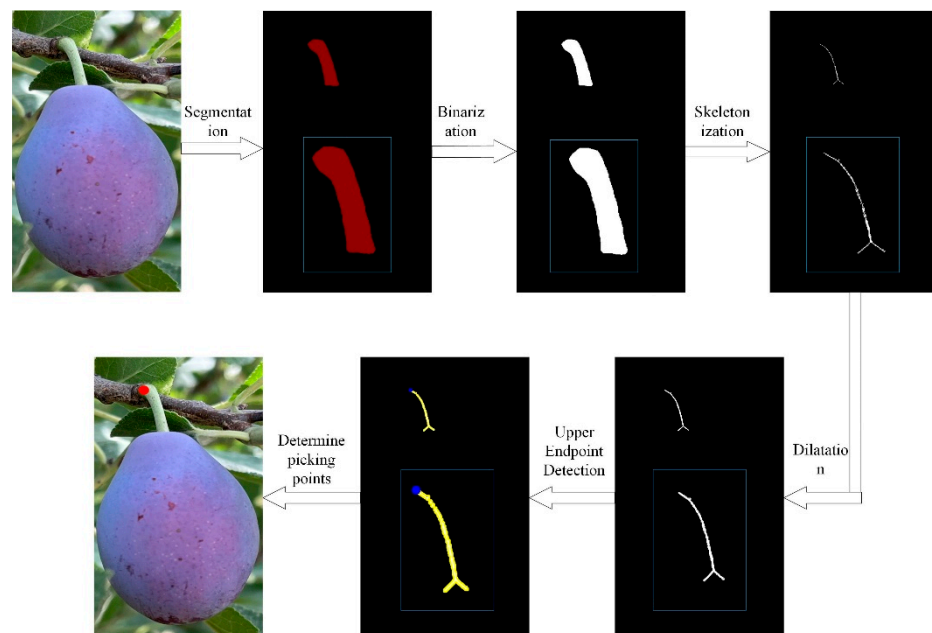


Figure 7. Flowchart of picking point localization.



The refinement algorithm employed in this study is the Zhang–Suen refinement algorithm [24], a binary image refinement technique based on iterative processing that removes unwanted pixel points while preserving only the skeleton points. In each iteration, the algorithm deletes pixels that satisfy specific conditions, progressively refining the target. This process continues through multiple iterations until all qualifying pixels are removed, at which point the algorithm concludes its operation, ultimately achieving the goal of target skeletonization.

## 4. Tests and Analysis

### 4.1. Experimental Environment

This study utilized a 64-bit Windows 10 operating system with 32 GB of RAM, an NVIDIA GeForce RTX 4080 graphics card, and a 13th Gen Intel® Core™ i7-13700KF processor. The experiments were conducted using the PyTorch 1.12.1 deep learning framework, CUDA version 12.0, and Python version 3.7. The training batch size was set to 8, with the Adam optimizer employed, an initial learning rate of 0.0005, and a weight decay parameter of 0.0001. The number of training iterations was 150.

### 4.2. Model Evaluation Indicators

To accurately assess the performance of the model, this study emphasizes the use of the mean intersection over union (MIoU), mean pixel accuracy (MPA), and model size as the evaluation metrics for semantic segmentation models.

The mean intersection over union (MIoU) represents the average of the ratio of the intersection and the merger of the results and the true values for each type of prediction, and it is calculated as follows:

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (5)$$

The mean pixel accuracy (mPA) represents the ratio of correctly predicted pixels to the total number of pixels and is calculated as follows:

$$\text{MPA} = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}} \quad (6)$$

where  $k+1$  represents the number of labels ( $k$ ) versus the total class of the background;  $P_{ii}$  denotes the number of pixels predicted correctly;  $P_{ij}$  denotes the number of pixels of class  $i$  predicted as class  $j$ ;  $P_{ji}$  denotes the number of pixels of class  $j$  predicted as class  $i$ .

### 4.3. Comparative Analysis of Test Results

#### 4.3.1. Comparative Tests of Attention Mechanisms

To investigate the influence of various attention mechanisms on model performance, this study performed comparative experiments by substituting the CBAM attention mechanism with three alternative mechanisms, SE (Squeeze-and-Excitation Networks) [25], ECA (Efficient Channel Attention Network) [26], and NAM (Normalization-based Attention Module) [27], within the context of the enhanced DeepLabv3+ model. The experimental results are presented in Table 1. As demonstrated in Table 1, the implementation of four distinct attention mechanisms—ECA, NAM, SE, and CBAM—resulted in an improvement in the model's MIoU by 0.69, 0.83, 1.13, and 1.31 percentage points, respectively, in comparison to the original DeepLabv3+ model. Concurrently, the model's MPA showed enhancements of 0.3, 0.33, 1.03, and 1.38 percentage points, respectively. Notably, the model incorporating the CBAM attention mechanism yielded the most favorable performance. This phenomenon can be attributed to the CBAM mechanism's integration of channel and

spatial attention, which facilitates the dual extraction of input features. Consequently, the model is better positioned to focus on critical channel and spatial location features, thereby significantly enhancing its segmentation accuracy.

**Table 1.** Comparative trials of different attention mechanisms.

Network Models	Backbone Network	Attention Module	MIoU/%	MPA/%
DeepLabv3+	Xception	NONE	84.82	91.54
DeepLabv3+	MobileNetv2	ECA	85.51	91.84
DeepLabv3+	MobileNetv2	NAM	85.65	91.87
DeepLabv3+	MobileNetv2	SE	85.95	92.57
DeepLabv3+	MobileNetv2	CBAM	86.13	92.92

#### 4.3.2. Comparison of Ablation Tests

To thoroughly investigate the impacts of the three proposed enhancement strategies on the model presented in this paper, four sets of ablation experiments were designed employing the control variable method; the experimental results are summarized in Table 2. As illustrated in Table 2, the adoption of the improved MobileNetv2 as the backbone network resulted in a marginal decline in segmentation accuracy due to model lightweighting, with the MIoU and MPA decreasing by 0.18 and 0.75 percentage points, respectively, when compared to the original DeepLabv3+ model. Following the incorporation of the CBAM attention mechanism in Experiment 2, the MIoU and MPA of the model increased by 0.32 and 1.2 percentage points, respectively, indicating that the integration of the CBAM attention mechanism significantly enhanced the model's segmentation accuracy. In Experiment 4, the incorporation of the DenseASPP module resulted in enhancements of 1.17 and 0.93 percentage points in the model's MIoU and MPA, respectively, when compared to Experiment 3. This indicates that the addition of the DenseASPP module improved the model's ability to focus on the details of the new plum fruit stalks, thereby significantly boosting its segmentation accuracy. When compared to the original DeepLabv3+ model, the simultaneous introduction of the three enhancement strategies led to improvements of 1.31 and 1.38 percentage points in the model's MIoU and MPA, respectively. This finding suggests that all three enhancement strategies exert positive effects on the model's performance.

**Table 2.** Ablation test.

Test	MobileNetv2	CBAM	DenseASPP	MIoU/%	MPA/%
1	×	×	×	84.82	91.54
2	✓	×	×	84.64	90.79
3	✓	✓	×	84.96	91.99
4	✓	✓	✓	86.13	92.92

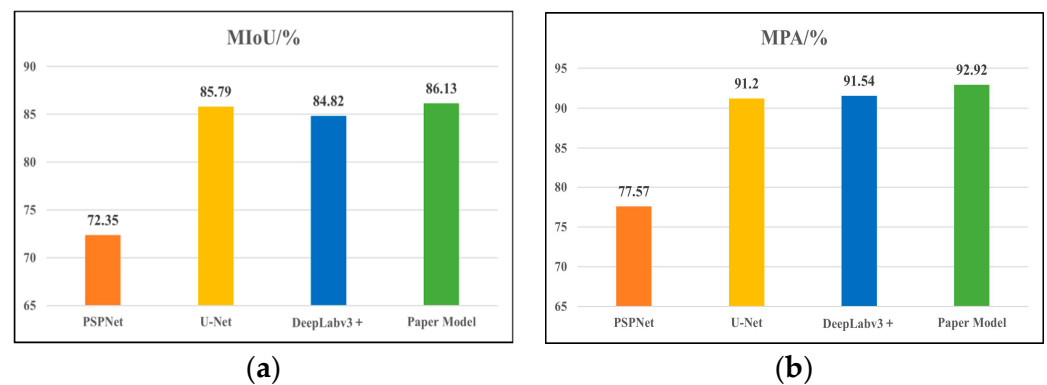
#### 4.3.3. Comparison Experiments of Different Segmentation Models

To effectively demonstrate the superiority of the improved model presented in this study, a comparative analysis was conducted against mainstream semantic segmentation models, namely PSPNet and U-Net, utilizing the same dataset. The experimental results, as illustrated in Table 3, along with the visualization results shown in Figure 8, indicate that compared to those of the PSPNet [28], U-Net [29], and DeepLabv3+ models, the MIoU of the proposed model was enhanced by 13.78, 0.34, and 1.31 percentage points, respectively, while the MPA showed improvements of 15.35, 1.72, and 1.38 percentage points. The segmentation performance of the proposed model was significantly superior to that of the other three models. In terms of model lightweighting, the proposed model had a size of only 59.6 MB, representing a reduction of 118.4, 107.4, and 149.4 MB, respectively, compared to the other three models. This model is particularly suitable for deployment on mobile devices. In summary, when compared to the PSPNet, U-Net, and DeepLabv3+ models,

the proposed model exhibited the highest MIoU and MPA, along with the lowest model size. It achieved high segmentation accuracy while maintaining lightweight characteristics. This paper proposes an enhancement of the DeepLabv3+ model for new plum fruit stalk recognition, referencing the advantages noted in the literature [12], which highlights the capability of maintaining high segmentation accuracy alongside a reduced model size. Consequently, the proposed model is more suitable for application in embedded removable devices. Unlike the approaches discussed in the literature [14], which have limitations due to the slim nature of new plum fruit stalks and the similar coloration of the leaves, the introduction of the DenseASPP module in our model significantly enhances the focus on the intricate details of segmentation specific to new plum fruit stalks, thereby fulfilling the requirements for accurate segmentation of new plum fruit stalks in the complex orchard environment.

**Table 3.** Comparative experiments on different models.

Network Models	Backbone Network	MIoU/%	MPA/%	Model Size/MB
PSPNet	Resnet50	72.35	77.57	178
U-Net	Resnet50	85.79	91.20	167
DeepLabv3+	Xception	84.82	91.54	209
Paper Model	MobileNetv2	86.13	92.92	59.6



**Figure 8.** Comparison chart for visualization of test results. (a) Test comparison MIoU chart; (b) Test comparison MPA chart.

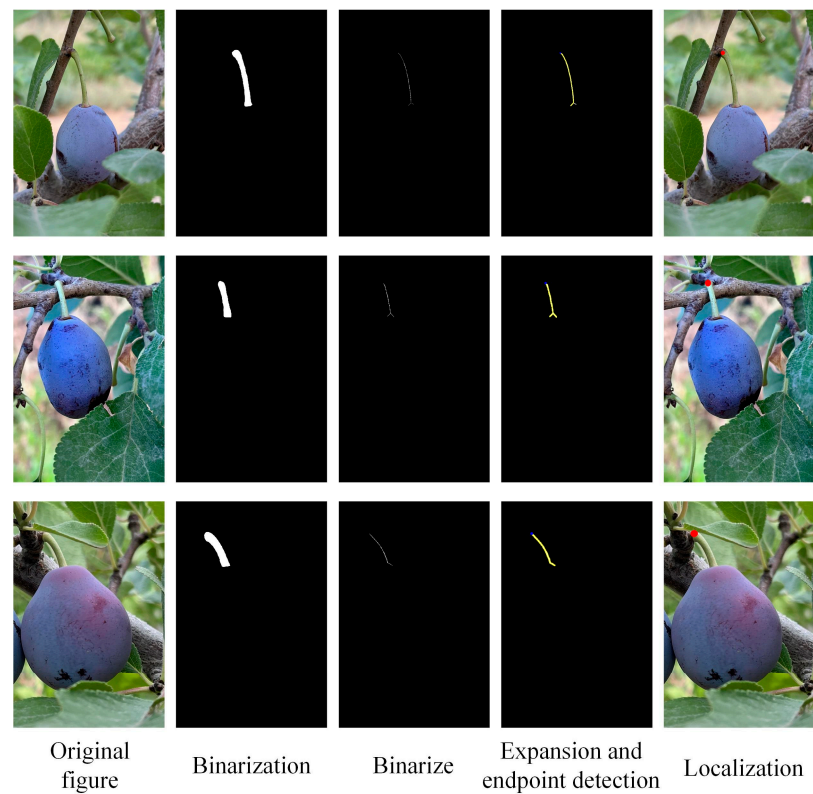
To facilitate a comparison of the segmentation performance across the four models, Figure 9 illustrates their respective effects on the test set, with the yellow box highlighting a magnified view of the regions exhibiting segmentation deficiencies. The yellow box specifically indicates the magnified region where segmentation inaccuracies occur. As illustrated in Figure 9, the U-Net model demonstrates superior segmentation of fruit stalks in unobstructed scenarios; however, it exhibits segmentation discontinuities and inaccuracies in edge delineation when faced with leaf occlusion. Both the PSPNet and original DeepLabv3+ models exhibit insufficient attention to the edge details of fruit stalks, resulting in mis-segmentation and jagged edge delineation. In comparison to the other three models, the enhanced DeepLabv3+ model achieves smoother segmentation of fruit stalk edges and minimizes omissions and misclassifications, thereby yielding the most favorable segmentation outcomes.



**Figure 9.** Segmentation effects of different models.

#### 4.3.4. Positioning Tests

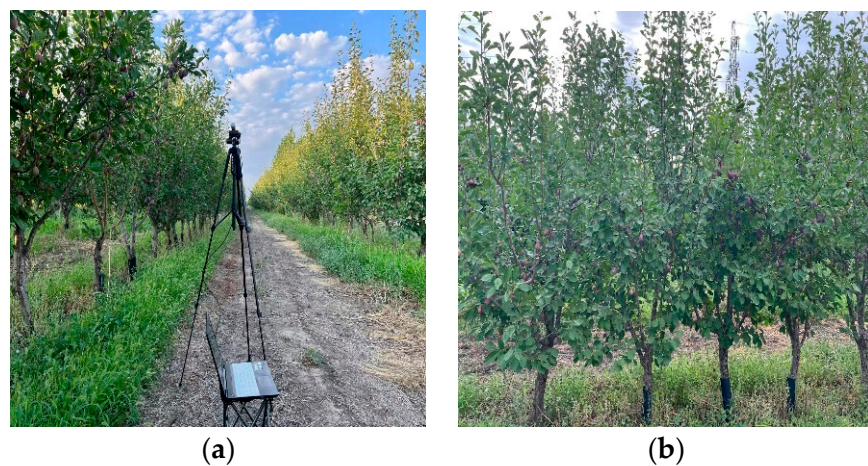
To further validate the reliability of the picking point positioning methodology introduced in this study, a selection of 73 images was drawn from the test dataset to conduct a picking point positioning experiment. The selected images encompassed 81 ripe new prune stalks, of which 72 samples were successfully located, while 9 samples were not successfully identified. The primary cause of the positioning failures was a significant influence of light on the fruit stems, coupled with an extensive area covered by branches and leaves. This ultimately led to inaccurate segmentation of the fruit stems and, consequently, imprecise positioning. To address the positioning failures arising from the influence of light and the substantial area of branches and leaves, implementing shading panels on the image acquisition equipment and adjusting the acquisition angle are proposed as viable solutions. Based on the experiments, the success rate of the newly developed plum picking positioning method reached 88.8%, with the positioning outcomes illustrated in Figure 10. In comparison to the literature [12,13], the newly developed plum picking positioning method not only guarantees accurate positioning within actual orchard environments but also is relatively straightforward to operate, thereby circumventing issues such as prolonged positioning times and inaccuracies arising from a complex positioning process. Consequently, the methodology delineated in this study exhibits robustness.



**Figure 10.** Successful mapping of picking point localization.

#### 4.3.5. Field Segmentation and Localization Experiments in Orchards

To further assess the feasibility of the segmentation and localization algorithm, we conducted a field test at a new plum plantation in Tsabchal County, Ili Kazakh Autonomous Prefecture, in August 2024. The test site is depicted in Figure 11. The experiment was conducted from 8:00 to 14:00 under sunny weather conditions, utilizing equipment that included an iPhone 13 for image capture, a camera tripod, a notebook, and a mobile control terminal. The experimental setup focused on the single-crop cultivation of new plums. Initially, images of the new plum fruit stalks were captured using the image acquisition equipment. Subsequently, these images were transmitted to the mobile control terminal for processing. The enhanced DeepLabv3+ model was employed to segment the fruit stalks, followed by the application of the endpoint localization method proposed in this paper to determine the picking point, yielding pixel coordinate outputs for the identified locations.

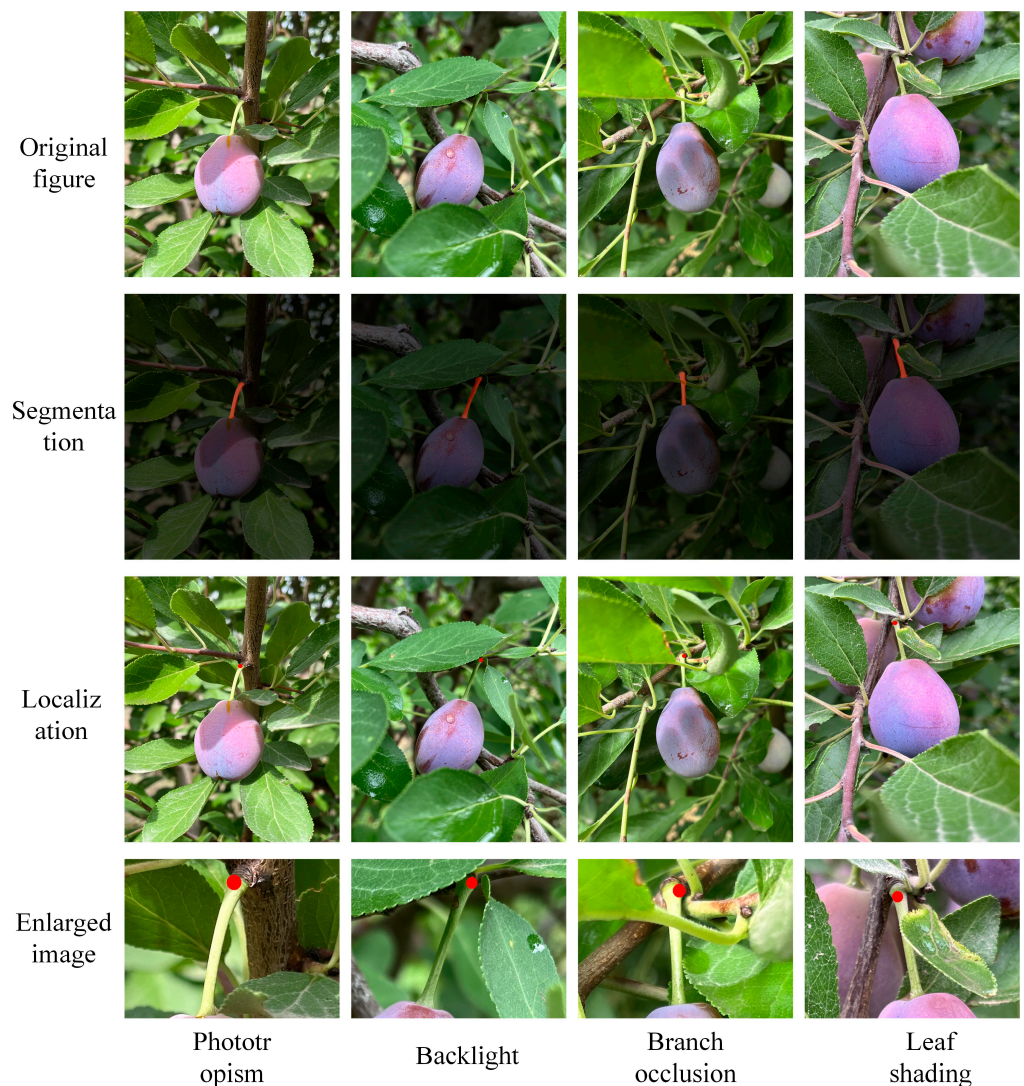


**Figure 11.** Field shooting test map. (a) Field experiment; (b) Test object.

A collection of images depicting the new plum fruit stalks was tested to select representative samples for segmentation and localization, with results illustrated in Figure 12. Under varying lighting conditions—including front light, backlight, and occlusion by branches and leaves—the improved DeepLabv3+ model successfully segmented the new plum fruit stalks with high accuracy. Following accurate segmentation, the endpoint localization method was employed to determine the picking point. In diverse orchard environments, the picking point was consistently localized at the upper end of the new plum stalk, with corresponding pixel coordinate values presented in Table 4. The field test validated the feasibility of the proposed segmentation and picking point localization method for new plum fruit stalks, simultaneously providing essential technical support for the subsequent development of new-plum-picking robots.

**Table 4.** Sample pixel coordinate points.

Serial Number	Test Environment	Pixel Coordinate Point
Sample 1	Phototropism	(1011, 918)
Sample 2	Backlight	(1303, 889)
Sample 3	Branch occlusion	(1049, 862)
Sample 4	Leaf shading	(719, 450)



**Figure 12.** Field segmentation positioning map.

## 5. Conclusions

Addressing the challenges of recognizing new plum stems and locating picking points within orchard environments, this study proposes an enhanced DeepLabV3+ network model designed for the segmentation and identification of fruit stems. The principal conclusions drawn from this research are as follows:

This study proposes a new plum fruit stalk recognition method based on an improved DeepLabV3+ framework. The method employs MobileNetv2 as the backbone network to minimize the model weight and incorporates the CBAM attention mechanism to enhance the model's capacity for extracting critical feature information from the fruit stalks. Furthermore, the introduction of the DenseASPP module improves the model's focus on edge details, thereby mitigating the occurrence of missed segmentation. The mean intersection over union (MIoU) of this model is 86.13%, while the mean pixel accuracy (MPA) stands at 92.92%, representing increases of 1.31 and 1.38 percentage points compared to the original model, respectively.

In comparative tests conducted on the same dataset, the model presented in this paper demonstrated significant improvements over the PSPNet and U-Net models, achieving increases in the mean intersection over union (MIoU) of 13.78 and 0.34 percentage points, as well as enhancements in the mean pixel accuracy (MPA) of 15.35 and 1.72 percentage points, respectively. Furthermore, this model exhibited the smallest size and the best overall performance, effectively meeting the requirements for real-time segmentation of new plum fruit stalks in orchard environments.

In this study, fruit stem thinning and endpoint detection algorithms were employed to accurately determine the final picking points of new plum stems. Additionally, the segmentation of fruit stem images facilitated binarization and skeletonization operations to extract the skeletal structure of the fruit stem, culminating in the application of the endpoint detection algorithm to finalize the identification of the picking point. The experimental results indicated that the success rate of locating the picking point was 88.8%, thereby offering valuable technical support for subsequent robotic-arm picking operations.

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

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 52465055; the Beijing Natural Science Foundation Project under Grant 6244056; and the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2023D01C189.

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

**Data Availability Statement:** The data generated in this study can be obtained from the author upon request. These data are not publicly accessible due to privacy and ethical considerations.

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

## References

1. Feng, Q.; Wang, X.; Wang, G.; Li, Z. Design and test of tomatoes harvesting robot. In Proceedings of the IEEE International Conference on Information and Automation, Lijiang, China, 8–10 August 2015; pp. 949–952.
2. Yao, Z.; Zhao, C.; Zhang, T. Agricultural machinery automatic navigation technology. *iScience* **2023**, *27*, 108714. [[CrossRef](#)] [[PubMed](#)]
3. Mishra, A.M.; Harnal, S.; Gautam, V.; Tiwari, R.; Upadhyay, S. Weed density estimation in soya bean crop using deep convolutional neural networks in smart agriculture. *J. Plant Dis. Prot.* **2022**, *129*, 593–604. [[CrossRef](#)]
4. Ionica, M.E.; Nour, V.; Trandafir, I.; Cosmulescu, S.; Botu, M. Physical and chemical properties of some European plum cultivars (*Prunus domestica* L.). *Not. Bot. Horti Agrobot. Cluj-Napoca* **2013**, *41*, 499–503. [[CrossRef](#)]
5. Wang, Z.; Wang, J.; Yang, K.; Wang, L.; Su, F.; Chen, X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. *Comput. Geosci.* **2022**, *158*, 104969. [[CrossRef](#)]

6. Kaur, P.; Harnal, S.; Tiwari, R.; Upadhyay, S.; Bhatia, S.; Mashat, A.; Alabdali, A.M. Recognition of leaf disease using hybrid convolutional neural network by applying feature reduction. *Sensors* **2022**, *22*, 575. [[CrossRef](#)] [[PubMed](#)]
7. Bac, C.W.; Hemming, J.; Van Henten, E.J. Stem localization of sweet-pepper plants using the support wire as a visual cue. *Comput. Electron. Agric.* **2014**, *105*, 111–120. [[CrossRef](#)]
8. Xiong, J.; Lin, R.; Liu, Z.; He, Z.; Tang, L.; Yang, Z.; Zou, X. The recognition of litchi clusters and the calculation of picking point in a nocturnal natural environment. *Biosyst. Eng.* **2018**, *166*, 44–57. [[CrossRef](#)]
9. Ji, C.; Zhang, J.; Yuan, T.; Li, W. Research on key technology of truss tomato harvesting robot in greenhouse. *Appl. Mech. Mater.* **2014**, *442*, 480–486. [[CrossRef](#)]
10. Luo, L.; Tang, Y.; Lu, Q.; Chen, X.; Zhang, P.; Zou, X. A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard. *Comput. Ind.* **2018**, *99*, 130–139. [[CrossRef](#)]
11. Yu, Y.; Zhang, K.; Liu, H.; Yang, L.; Zhang, D. Real-time visual localization of the picking points for a ridge-planting strawberry harvesting robot. *IEEE Access* **2020**, *8*, 116556–116568. [[CrossRef](#)]
12. Peng, H.; Xue, C.; Shao, Y.; Chen, K.; Xiong, J.; Xie, H.; Zhang, L. Semantic segmentation of litchi branches using DeepLabV3+ model. *IEEE Access* **2020**, *8*, 164546–164555. [[CrossRef](#)]
13. Ning, Z.; Lou, L.; Liao, J.; Wen, H.; Wei, H.; Lu, Q. Recognition and the optimal picking point location of grape stems based on deep learning. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 222–229.
14. Rong, Q.; Hu, C.; Hu, X.; Xu, M. Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing. *Comput. Electron. Agric.* **2023**, *210*, 107923. [[CrossRef](#)]
15. Yan, C.; Chen, Z.; Li, Z.; Liu, R.; Li, Y.; Xiao, H.; Xie, B. Tea sprout picking point identification based on improved DeepLabV3+. *Agriculture* **2022**, *12*, 1594. [[CrossRef](#)]
16. Wu, L.; Su, L.; Jia, G.; Ma, Y.; Li, B.; He, S. Image Segmentation of Potato Roots Using an Improved DeepLabv3+ Network. *Trans. Chin. Soc. Agric. Eng.* **2023**, *39*, 134–144.
17. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
18. Zhu, Y.; Liu, S.; Wu, X.; Gao, L.; Xu, Y. Multi-class segmentation of navel orange surface defects based on improved DeepLabv3+. *J. Agric. Eng.* **2024**, *55*. [[CrossRef](#)]
19. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1251–1258.
20. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
21. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
22. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
23. *DB65/T 4475-2021*; Quality Grading of *Prunus Domestica*. Market Supervision Administration of Xinjiang Uygur Autonomous Region: Urumqi, China, 2021.
24. Zhang, T.Y.; Suen, C.Y. A fast parallel algorithm for thinning digital patterns. *Commun. ACM* **1984**, *27*, 236–239. [[CrossRef](#)]
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
26. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
27. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419.
28. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
29. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.; Zhang, X.; Huang, X. Road detection and centerline extraction via deep recurrent convolutional neural network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [[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.