applied sciences

MDPI

Article

A Multi-Target Identification and Positioning System Method for Tomato Plants Based on VGG16-UNet Model

Xiaojing Li [1,2,3], Jiandong Fang [1,2,3,*] and Yvdong Zhao [2,3]

1 College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010080, China
2 Inner Mongolia Key Laboratory of Perceptual Technology and Intelligent Systems, Hohhot 010080, China
3 Collaborative Innovation Center of Perception Technology in Intelligent Agriculture and Animal Husbandry, Inner Mongolia Autonomous Region, Hohhot 010080, China
* Correspondence: fangjd@imut.edu.cn

Abstract: The axillary buds that grow between the main and lateral branches of tomato plants waste nutrients and lead to a decrease in yield, necessitating regular removal. Currently, these buds are removed manually, which requires substantial manpower and incurs high production costs, particularly on a large scale. Replacing manual labor with robots can lead to cost reduction. However, a critical challenge is the accurate multi-target identification of tomato plants and precise positioning for axillary bud removal. Therefore, this paper proposes a multi-target identification and localization method for tomato plants based on the VGG16-UNet model. The average intersection and pixel accuracies of the VGG16-UNet model after introducing the pretrained weights were 85.33% and 92.47%, respectively, which were 5.02% and 4.08% higher than those of the VGG16-UNet without pretrained weights, achieving the identification of main branches, side branches, and axillary bud regions. Then, based on the multi-objective segmentation of the tomato plants in the VGG16-UNet model, the regions of the axillary buds in the tomato plants were identified by HSV color space conversion and color threshold range selection. Morphological dilation and erosion operations were used to remove noise and connect adjacent regions of the same target. The endpoints and centroids of the axillary buds were identified using the feature point extraction algorithm. The left and right positions of the axillary buds were judged by the relationship between the position of the axillary bud centroid and the position of the main branch. Finally, the coordinate parameters of the axillary bud removal points were calculated using the feature points to determine the relationship between the position of the axillary bud and the position of the branch. Experimental results showed that the average accuracy of the axillary bud pruning point recognition was 85.5%.

Keywords: tomato plant; axillary bud; image recognition; object detection; VGG16-UNet; removal point localization

check for updates

Citation: Li, X.; Fang, J.; Zhao, Y. A Multi-Target Identification and Positioning System Method for Tomato Plants Based on VGG16-UNet Model. *Appl. Sci.* **2024**, *14*, 2804. https://doi.org/10.3390/app14072804

Academic Editors: Yujin Lim and Hideyuki Takahashi

Received: 4 March 2024
Revised: 24 March 2024
Accepted: 25 March 2024
Published: 27 March 2024

1. Introduction

With the advancement of agricultural science and technology, improving the efficiency of monitoring and managing the growth status of crops is crucial for the agricultural industry. Among the world's most widely grown vegetables, tomatoes stand out because the growth of their shoots has a significant impact on both yield and quality, particularly the axillary buds that develop between the main stem and the base of the lateral branches. However, the current method of removing axillary buds relies on manual labor, resulting in significant labor requirements and high production costs, especially in large-scale cultivation [1,2]. The wide row spacing typical of large-scale tomato production makes the automation of axillary bud removal feasible [3]. Therefore, the development of a tomato axillary bud remover robot to replace manual labor is not only feasible but imperative. To achieve automated tomato axillary bud pruning, the accurate detection of tomato main branches, side branches, and axillary buds and the identification of pruning points are of paramount importance.

Currently, scholars have conducted relevant research on the automatic pruning of plant branches and fruit and vegetable harvesting. In 2017, the Dutch company Priva released a tomato branch and leaf-pruning robot, which can achieve the automatic pruning of tomato branches and leaves [4]. Ning et al. [5] improved Mask R-CNN to identify and segment grape stems, achieving an average recognition accuracy of 88% for grape stems. Yan et al. [6] used the K-means clustering method to identify the branches of Lycium barbarum, overcoming the problem of numerous interferences in the identification of wolfberry branches in natural environments. Peng et al. [7] used DeepLabv3+ to segment litchi branches and obtained an average intersection over union ratio of 76.5% for model segmentation. Peng et al. [8] proposed an improved deep learning model, RDf-DeepLabV3+, to address the semantic segmentation of lychee branches in an orchard environment. The main improvement of this model is the introduction of a new backbone network, called ResDense, which integrates the ResNet and DenseNet networks and replaces the original backbone network of DeepLabV3+. In addition, focal loss is used to replace the original cross-entropy loss function. Experimental results show that the RDf-DeepLabV3+ model outperforms all comparison models. Specifically, the mean intersection over union (mIoU) of the RDf-DeepLabV3+ model is 0.848, 0.811, and 0.770 for images of simple, moderate, and complex levels, respectively. Furthermore, the training and testing speeds of this model are approximately 7.7% faster than Xception. This research provides a reliable deep learning solution for the accurate segmentation of lychee branches in orchard environments, which is of great practical importance for robotic lychee harvesting. Palacios et al. [9] combined VGG19 with Segnet for the detection and segmentation of grape berries, achieving F1 scores of 0.93 and 0.73, respectively. Afonso et al. [10] built a Mask R-CNN-based model for the identification of ripe and unripe tomatoes, with recognition accuracies of 95% and 94%, respectively. Wei et al. [11] used the DA2-YOLOv4 model to identify hedges with a recognition speed of 83.1 frames per second and an average accuracy of 98.5%. Ma et al. [12] used an improved Mask R-CNN model to segment rice stalk contamination, achieving a segmentation and detection accuracy of 91.12% with an average processing time of 3.57 s. Liang et al. [13] used a fusion of YOLOv3 and UNet to detect and segment lychee and fruit stalks in a nocturnal environment, demonstrating high accuracy and robustness. Jia et al. [14] proposed an improved apple-picking robot vision detector model based on a modified Mask R-CNN. They combined ResNet and DenseNet as the backbone network for feature extraction, aiming to reduce input parameters while retaining valuable features for accurate detection. The results show that the model achieved a precision of 97.31% and a recall of 95.70%, meeting the requirements of the practical apple-picking robot vision system in orchard environments. Liang et al. [15] segmented tomato plants using the improved Mask R-CNN model and achieved the positioning of tomato lateral branch pruning points through mask merging and conditional constraints. The average accuracy of pruning point identification was 82.9%.

The above studies indicate some progress in the field of plant pruning and fruit and vegetable harvesting. However, research into the identification of tomato axillary pruning points is limited, with traditional image processing methods being predominantly used. Traditional image processing methods and convolutional neural network detection techniques are common approaches to target recognition. Due to the similar colors of tomato lateral branches, main branches, and axillary buds, traditional threshold segmentation methods perform adequately in certain scenarios. However, they are sensitive to variations in plant state and growth environment, resulting in limited detection capabilities in complex and changing growth environments. In contrast, deep learning models can effectively detect the main branch, lateral branch, and axillary bud targets of tomato plants by leveraging multi-dimensional features such as color, shape, and texture. These models demonstrate adaptability to different scenarios and changes in plant status, thereby improving the ability to detect multiple targets of tomato plants in dynamic growth environments.

To address the challenge of effectively segmenting axillary buds from a background of similar colors using traditional image processing methods and to ensure robust recognition

in dynamic environments, this paper proposes a methodology. We use the VGG16-UNet deep learning model to achieve the segmentation of tomato plant main branches, lateral branches, and axillary buds. In addition, we use techniques such as HSV color space conversion and threshold segmentation to achieve fine segmentation and separation of the main branches, lateral branches, and axillary buds. Feature points are then extracted from the segmented regions. Subsequently, the axillary bud growth position is determined based on these feature points, and the tomato axillary bud pruning point is calculated using positional relationships. This approach aims to provide technical support for the identification of pruning points by tomato axillary bud-picking robots.

The main contributions and innovations of this study are outlined as follows:

- Proposal of a multi-target segmentation method for tomato plants based on the VGG16-UNet model. This method effectively performs pixel-level segmentation of the main stem, lateral branches, and axillary buds of tomato plants, providing a solid basis for calculating axillary bud removal points for automated agricultural robots.
- By combining the powerful VGG16 and ResNet models with the UNet architecture and using pretrained weights, the accuracy and training efficiency of the model for tomato plant image recognition is significantly improved. This fusion enhances feature extraction capabilities and improves model generalization, enabling better adaptation to complex agricultural scenarios.
- Development of a fine segmentation algorithm that combines HSV color space transformation and morphological operations. This algorithm effectively separates key structures of tomato plants from complex backgrounds, improving the accuracy of axillary bud positioning and providing reliable technical support for automated agricultural robots.
- The use of traditional semantic segmentation evaluation metrics alongside the introduction of pixel-length correspondence analysis and specific rule evaluations. This approach provides a new perspective for comprehensive model evaluation, ensuring the effectiveness and reliability of the model in practical applications.

The remainder of this paper is divided into five sections. Section 2 presents the methods and technical approach used in this study. Section 3 describes the composition of the dataset. Section 4 provides a detailed description of the models and methods used. Section 5 discusses the details of the experimental design, including parameters for model training, evaluation metrics for assessing model performance, and a comprehensive analysis of the experimental results. Finally, in Section 6 a comprehensive discussion of the experimental results is provided, along with suggestions for possible future research directions.

2. Research Methods

The end effector pruning operation of the axillary bud removal robot is illustrated in Figure 1 [15]. To ensure efficient pruning with the end effector, the process begins with the inspection of the tomato plant. This step is crucial as the robot must accurately recognize the position and structure of the plant in order to calculate the point of axillary bud removal for subsequent operations. Figure 2 outlines the technical path of this document.

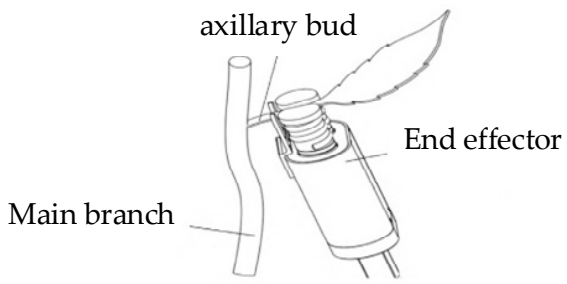


Figure 1. Pruning operation with end effector.

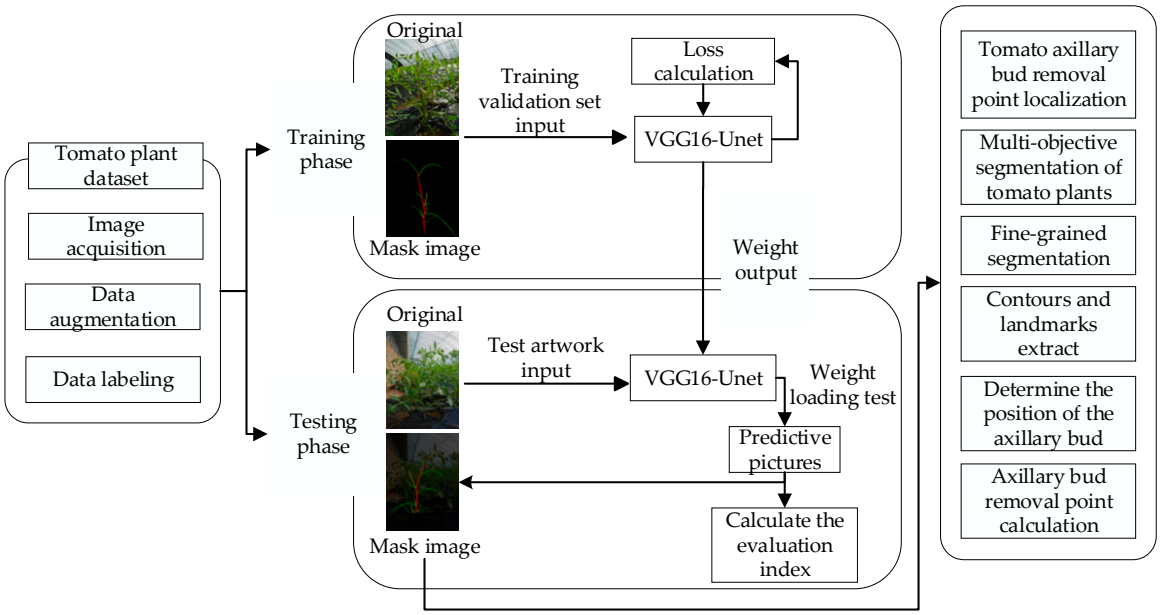


Figure 2. Technology roadmap.

The workflow starts by annotating the collected images and applying data augmentation techniques to construct a dataset tailored to complex real-world environments. Next, this dataset is fed into the VGG16-UNet network for training, aiming to develop a highly efficient prediction model. This model will then be used to accurately detect the main branches, lateral branches, and axillary bud areas of tomato plants in real-world scenarios.

Afterward, image processing techniques were used to finely segment the prediction results of the neural network, focusing particularly on the main branch and axillary bud areas. This involved contour extraction and landmark identification to achieve a more accurate delineation of the plant structure. Finally, by analyzing the positions of the feature points, we were able to determine the exact location of the axillary bud and calculate the point for its removal.

3. Dataset Production

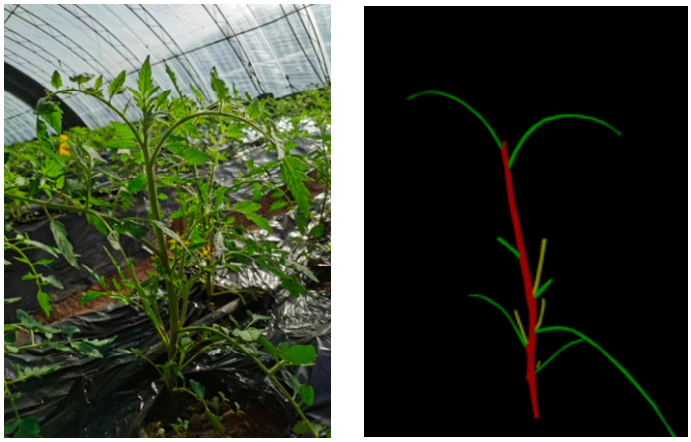
The quality and diversity of the dataset has a direct impact on the model's ability to learn features and generalize. This section first outlines the data collection and annotation processes. Next, color jittering data augmentation is used to simulate different lighting and environmental conditions, thereby enhancing the dataset and improving the model's adaptability to different complex scenes. Finally, the original and augmented images are merged into a balanced dataset, which is then partitioned into training, validation, and test sets for subsequent model training and evaluation in subsequent chapters.

3.1. Data Collection and Annotation

Based on the results of the survey, the removal of tomato axillary buds typically begins around the seedling stage of approximately 30 days, with the targeted buds being approximately 7 cm in length. Consequently, this study selected tomato seedlings aged around one month old as experimental subjects, with the aim of reducing the risk of disease due to the challenges associated with healing large wounds.

In this article, the image capture device has a resolution of 2400 × 1080, and the screen measures 6.57 inches. The image data were collected from Hohhot Green Union Planting Professional Cooperative, resulting in total of 541 images of tomato plants at the seedling stage in JPG format. Figure 3a shows the original image of a tomato plant featuring 7 tomato objects, including 1 main branch, 4 lateral branches, and 2 axillary buds. As the axillary bud is the target for pruning operations, with the main branch and lateral branch serving as auxiliary objects for axillary bud recognition, we used the version 3.16.7

of the Labelme tool to annotate the root areas of the axillary bud, main branch, and lateral branch point by point, generating a json file. This file was then converted to produce the corresponding labeled image. As shown in Figure 3b, the red area represents the main branch area, the green area represents the lateral branch area, the yellow area represents the axillary bud area, and the black area represents the background area.

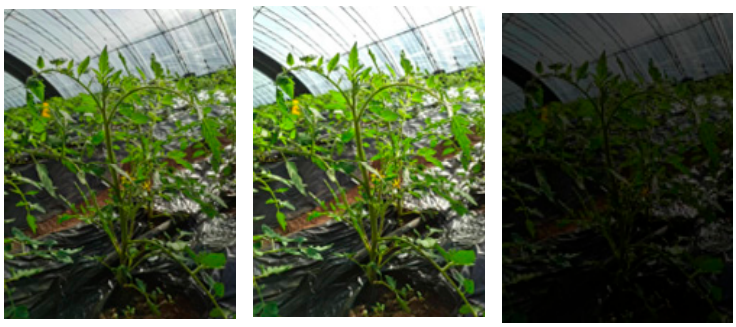


(a) original image (b) mask of marking

Figure 3. Marking the axillary bud, main branch, and lateral branch.

3.2. Data Augmentation

Due to the limited number of images collected, color jitter brightness enhancement was used to augment the dataset in order to increase the volume of training data and improve the model's ability to generalize. This method alters the brightness of the images through shading while ensuring that the parameters remain consistent with the natural environmental state to avoid image distortion. Specifically, it simulates both strong and weak lighting conditions by adjusting the brightness accordingly. After data enhancement, a total of 1082 enhanced images were obtained. Consequently, the final sample set comprised a total of 1623 images, including both the original and enhanced images. Following an 8:1:1 ratio, these images were divided into 1299 images for the training set, 162 images for the validation set, and 162 images for the test set. Sample images with partial data enhancement are shown in Figure 4.



Original Increase the brightness Reduce the brightness

Figure 4. Data augmentation.

4. Multi-Objective Segmentation and Localization Method for Tomato Plants

4.1. Based on VGG16-UNet Tomato Plant Segmentation Model

Due to its excellent segmentation performance and fast training speed on small sample datasets [16–19], the UNet network is adopted by this paper as the basic model for the segmentation task. The UNet architecture consists of an encoder and a decoder. The

encoder extracts features from the input image to obtain four layers of feature maps, while the decoder performs layer-by-layer upsampling through deconvolution operations, accurately locating features and fusing them with the corresponding feature maps obtained at each level by the encoder. At the same time, the VGG16 model demonstrates robust image feature extraction and analysis capabilities. The use of large datasets for pretraining through transfer learning can effectively improve model performance and generalization ability. Therefore, the encoder part of the UNet model is replaced by the first 13 convolutional layers of VGG16, resulting in the VGG16-UNet model, as shown in Figure 5.

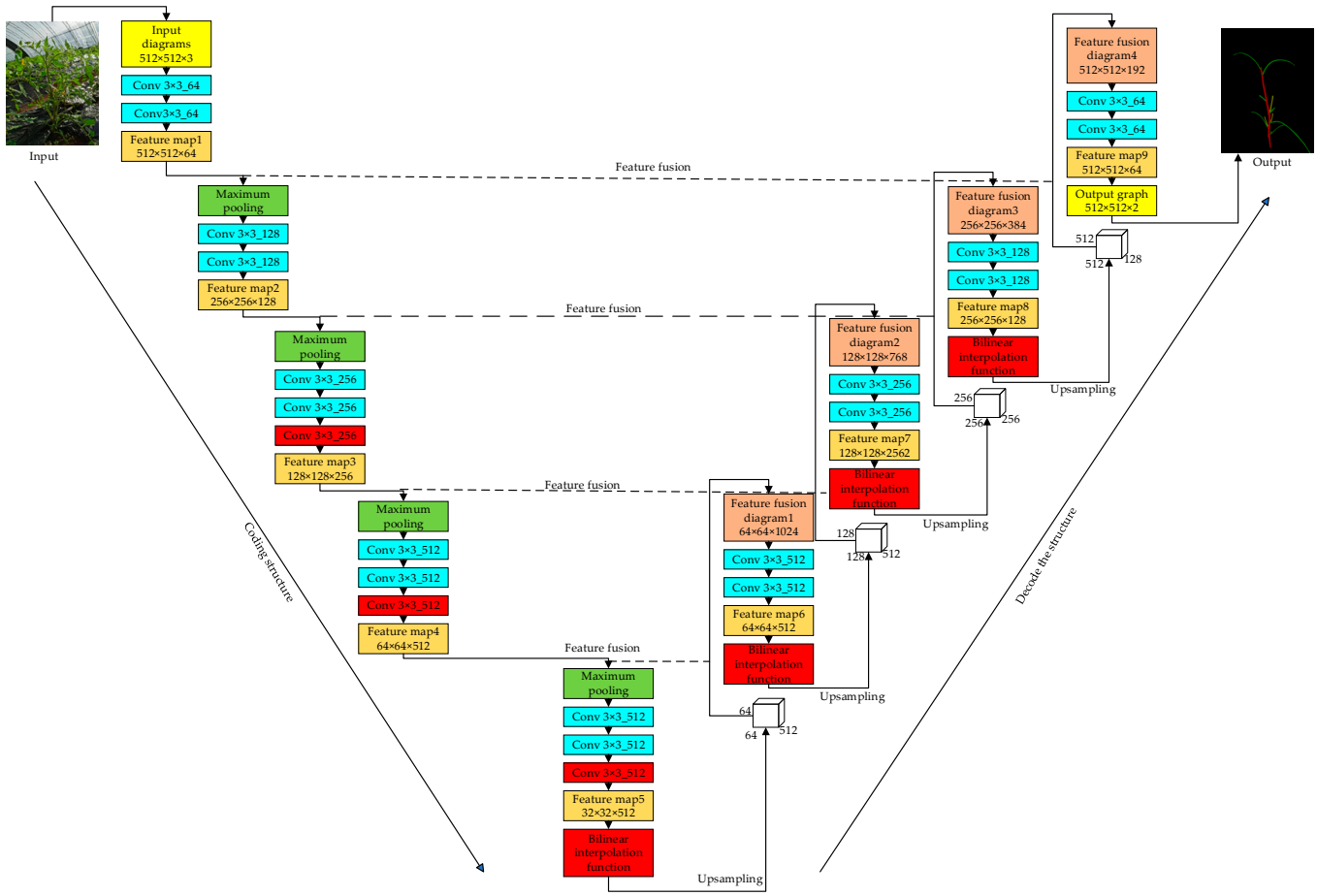


Figure 5. VGG16-UNet architecture.

In this paper, tomato plant segmentation is performed using the VGG16-UNet semantic segmentation model, and the parameter variations of each layer of the segmentation model are shown in Figure 5. The first 13 convolutional layers of VGG16 serve as the feature extraction network in the encoding structure. Each convolutional layer used a 3 × 3 kernel size, with the number of kernels being 64, 128, 256, and 512, respectively. The activation function used is ReLU. The encoder uses a stacking operation of convolution and pooling to perform downsampling, progressively enlarging the receptive field and compressing the image from an input size of 512 × 512 × 3 to a size of 32 × 32 × 512 after four rounds of downsampling, thereby condensing the features of the tomato plant images. In the corresponding decoding part, upsampling and feature fusion are used to achieve tomato plant segmentation. Upsampling is performed using bilinear interpolation to increase the size of the feature map, which speeds up the training of the tomato plant segmentation model. Through continuous upsampling and convolutional stacking, the decoder reconstructs the tomato plant map, restoring the size of the final output layer to 512 × 512. It then outputs the binary segmentation map of the tomato plant morphology. The VGG16-UNet semantic segmentation model uses four rounds of feature fusion to

perform plant segmentation. The convolutional feature map in the encoder is fused with the corresponding upsampled feature map in the decoder. Using Skip-Connection cascade fusion, shallow position information and deep semantic information of tomato plants are fused across multiple channels to achieve the pixel-level segmentation of tomato plants.

4.2. Location of Axillary Bud Removal Points of Tomato Plants

In this paper, the process of the axillary bud removal point localization method for tomato plants based on VGG16-UNet is shown in Figure 6.

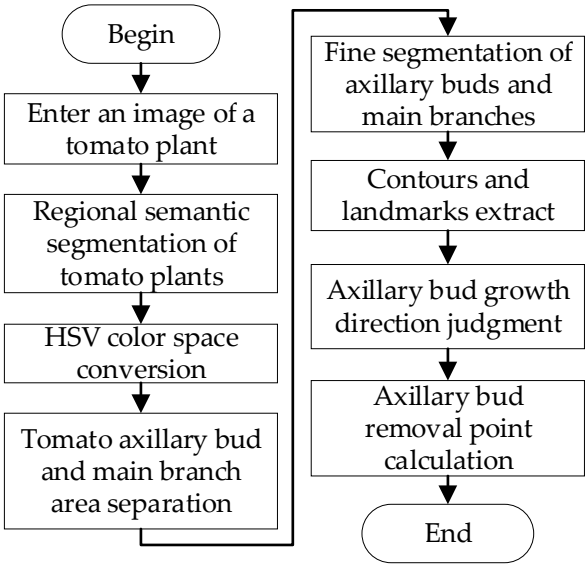


Figure 6. Tomato axillary bud removal point positioning process.

4.2.1. Fine Segmentation of Tomato Plants

In order to achieve an accurate segmentation of tomato plants, which is crucial for the location of removal points, this study used image processing methods following the coarse identification and segmentation of tomato plants based on VGG16-UNet. The steps involved are outlined below.

Step 1: Obtain the multi-target coarse segmentation result map of the tomato plant.

Step 2: Convert the RGB result image obtained from the coarse segmentation to HSV color space and separate the plant region from the background.

The HSV color space has three components: hue (H), saturation (S), and value (V). Hue represents a specific color, and using only hue can conveniently represent the distribution of colors in an image, making it simpler than the RGB model in directly describing colors [20]. After segmentation by the VGG16-UNet model, different targets of tomato plants have specific colors. By adjusting the threshold range of the HSV channels, specific color regions can be selected relatively easily without considering the complex combination of the red, green, and blue channels in the RGB color space and the weight of each component [21]. In addition, compared to the RGB model, the HSV model is more robust and less sensitive to changes in lighting conditions, which is particularly important in outdoor environments where tomato plants may be exposed to variations in sunlight. Therefore, in this experiment, the segmented mask images obtained from the VGG16-UNet semantic model are transformed from the RGB color space to the HSV color space for the subsequent extraction of the axillary regions and calculation of axillary removal points. The RGB images are separated into channels, and, after separation, each of the R, G, and B channels is normalized according to Equation (1) to obtain r, g, and b.

$$r = \frac{R}{255}, g = \frac{G}{255}, b = \frac{B}{255} \tag{1}$$

After the normalized channels (r, g, b) are obtained, they are utilized to calculate the corresponding HSV channels (H, S, V). The HSV channels are calculated as follows:

$$V = \max \tag{2}$$

$$S = \begin{cases} \frac{\max - \min}{\max}, \text{ if } \max \neq 0 \\ 0^\circ, \quad otherwise \end{cases} \tag{3}$$

$$H = \begin{cases} 0^\circ, & \text{if } \max = \min \\ 60^\circ \times \frac{(g-b)}{(\max - \min)} + 0^\circ, & \text{if} \max = randg \geq b \\ 60^\circ \times \frac{(g-b)}{(\max - \min)} + 360^\circ, & \text{if} \max = randg < b \\ 60^\circ \times \frac{(b-r)}{(\max - \min)} + 120^\circ, & \text{if} \max = g \\ 60^\circ \times \frac{(r-g)}{(\max - \min)} + 240^\circ, & \text{if} \max = b \end{cases} \tag{4}$$

The variables H, S, and V represent the values of the RGB image after conversion to the HSV color space model. Here, max corresponds to the maximum value among r, g, and b, while min corresponds to the minimum value among r, g, and b.

Step 3: Adjust the threshold range to achieve the separation of axillary buds of the tomato plants. The thresholds values for the three channels of H, S, and V are set to [20, 40], [100, 255], and [100, 255], respectively, to achieve axillary bud separation.

Step 4: Morphological manipulations

Using a 7×7 cross-shaped structural element, perform erosion and dilation operations on the binary image of the axillary bud region to remove interfering points and connect adjacent areas belonging to the same shoot, thereby achieving finer segmentation of tomato plants.

4.2.2. Axillary Bud Removal Point Calculation

Having achieved the accurate multi-target segmentation of tomato plants, the next critical step is to accurately calculate the removal points for the axillary buds. This process is critical for automated robotic pruning operations, as it directly affects pruning efficiency and future plant growth. To ensure healthy post-pruning plant growth and increased yield, the accurate localization of the axillary bud removal points requires consideration of both the morphological features of the axillary buds and their positional relationships to the main stem. In this section, we present the method for calculating the axillary bud removal points based on the segmentation results obtained from the VGG16-UNet model. This method includes contour detection, feature point extraction, axillary bud position determination, and precise removal point calculation. Through this series of accurate calculation steps, our aim is to provide the robot with accurate pruning guidance, thereby enabling efficient and non-destructive axillary bud pruning.

Step 1: Contour detection and feature point extraction

The contour of the axillary bud region was delineated, and the left and right endpoints of the axillary bud contour were determined, followed by the calculation of its centroid. The centroid, denoted as $\mathrm{p}(n_c, m_c)$, is calculated by the weighted average of the coordinates of each contour point, as described by the following formula:

$$n_c = \frac{M_{10}}{M_{00}} \tag{5}$$

$$m_c = \frac{M_{01}}{M_{00}} \tag{6}$$

$$M_{10} = \sum_x \sum_{\mathrm{y}} x \cdot I(x, y) \tag{7}$$

$$M_{01} = \sum_x \sum_{\mathrm{y}} \mathrm{y} \cdot I(x, y) \tag{8}$$

$$M_{00} = \sum_x \sum_{\mathrm{y}} I(x, y) \tag{9}$$

where n_c and m_c, respectively, represent the abscissa and ordinate of the axillary bud centroid, while c denotes the centroid index, where (c = 1, 2, 3...), indicates the pixel value at the center (x,y) of the region.

Step 2: Judging the left and right branches of the axillary buds

The vertical axis of each axillary bud center coordinate was extracted. Then, the contour points of the main branch were traversed to find the points whose vertical axis coincided with that of the axillary bud center. Subsequently, the horizontal axes of the two points were compared to determine the left and right branches of the axillary buds.

Step 3: Calculating the axillary bud removal point

According to agricultural standards, the optimum length of axillary bud stubble is approximately 1 cm, a criterion that aligns with the agronomic requirements [15]. Observations have shown that the points of axillary buds at around 30 days are predominantly in the range of 7 ± 1 cm. Therefore, in this study, the axillary bud removal was chosen near the end of the main branch and positioned at 1/4 distance from the center of the axillary bud. This choice ensures that the axillary bud removal point meets the pruning requirements.

The tomato plant has an axillary bud (Edge1) at its left end and another axillary bud (Edge2) at its right end. The coordinates (s_1, t_1), (s_2, t_2), and (n_1, m_1) represent the pixel values of the left endpoint, the right endpoint, and the centroid of Edge1, respectively. Similarly, (s_4, t_4), (s_5, t_5), and (n_2, m_2) denote the pixel values of the left endpoint, right endpoint, and centroid of Edge2, as shown in Figure 7.

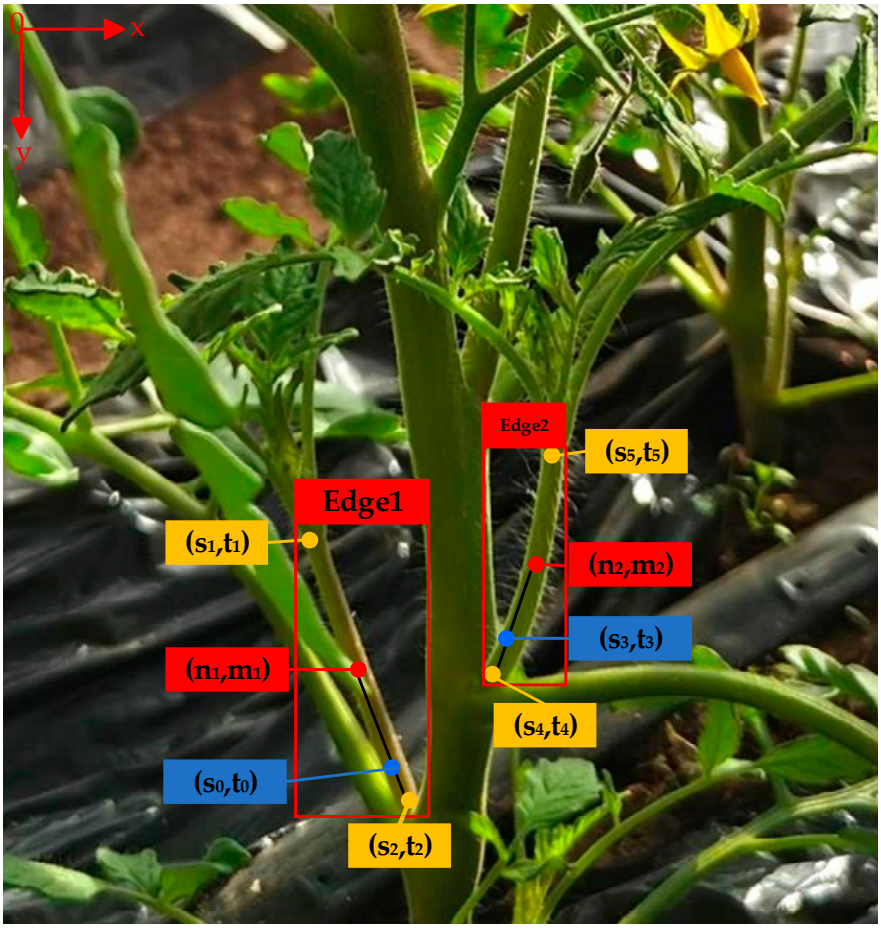


Figure 7. Tomato axillary bud removal point positioning.

The pruning point coordinates for the axillary bud (Edge1) located at the left end of the main branch mask are calculated as follows:

$$\begin{cases} s_0 = \frac{3s_2+n_1}{4} \\ t_0 = \frac{3t_2+m_1}{4} \end{cases} \tag{10}$$

where s_0 and t_0 are the coordinates of the trim point of Edge1.

The formula for calculating the coordinates of the pruning point when the axillary bud (Edge1) is at the right end of the main branch mask is as follows:

$$\begin{cases} s_3 = \frac{3s_4+n_2}{4} \\ t_3 = \frac{3t_4+m_2}{4} \end{cases} \tag{11}$$

where s_3 and t_3 are the coordinates of the trimming point of Edge2.

4.3. Experimental Workflow

The identification of tomato plants and the localization of axillary bud removal points can be summarized as Algorithm 1.

Algorithm 1: Multi-target identification and localization of tomato plants based on the VGG16-UNet model

Input:
Tomato plant image data
Output:
Multi-target mask map of a tomato plant;
Coordinates of tomato axillary bud removal points;
for each t ranging from 1 to the last tomato plant image N
1. Image pre-processing, standardization, resizing;
2. Implement encoder feature extraction based on the left half of Figure 5;
3. Implement a featured decoder based on the right half of Figure 5, perform pixel-level segmentation of the tomato plant image;
4. Obtain the masked images for segmentation of tomato plant main stem, lateral branches, and axillary buds;
5. Convert the masked image to HSV color space;
6. The yellow color range was extracted to obtain a binary graph containing only the axillary bud region;
7. Based on 6, morphological operations were used to obtain the binary graph after fine segmentation of axillary buds;
8. The Canny algorithm was used to find the contour of the binary image;
9. **For** each contour region in the binary segmented image:
a. left_point, right_point, centre_point = CalculateContourPoints(contour);
b. Determine whether the axillary bud (contour) is on the left or right side;
If on the left:
The take point for the axillary bud is 1/4 from the right endpoint to the midpoint;
Else:
The axillary bud removal point is 1/4th from the left endpoint to the midpoint;
c. Predicted axillary bud removal points in tomato plant images;
End for

According to Algorithm 1, the steps for the multi-target segmentation and localization of tomato plants using the example image in Figure 8a are detailed as follows.

In step 1, the image is pre-processed by denoising, adjusting the brightness and contrast, and resizing it to obtain a standardized image ready for input into the VGG16-UNet model. In step 2, the pre-processed image is fed into the encoder section of the VGG16-UNet model to extract image features. In step 3, a pixel-level semantic segmentation of the tomato plant image is performed by the decoder section of the model to obtain a segmentation mask image (Figure 8b), where different colors represent different regions

such as main branches, lateral branches, and axillary buds. Next, in step 5, the segmentation mask image is converted from the RGB color space to the HSV color space for more accurate color information processing (Figure 8c). In step 6, based on the yellow threshold in the HSV color space, a binary mask containing only the axillary bud region is generated (Figure 8d). In step 7, morphological operations such as erosion and dilation are applied to the binary mask image to remove noise and more precisely define the axillary bud region, resulting in a refined binary image (Figure 8e). Using the Canny algorithm (step 8), contours are detected in the refined binary image, and key points for each contour, including the left endpoint, right endpoint, and centroid, are calculated (step 9a). Based on these points, the position of the axillary buds is determined, and the removal points are calculated (step 9b), which are then marked in Figure 8f. Finally, in step 9c, the calculated axillary bud removal points are marked on the original image, resulting in the final image (Figure 8g), which will guide automated agricultural robots in precise axillary bud pruning.

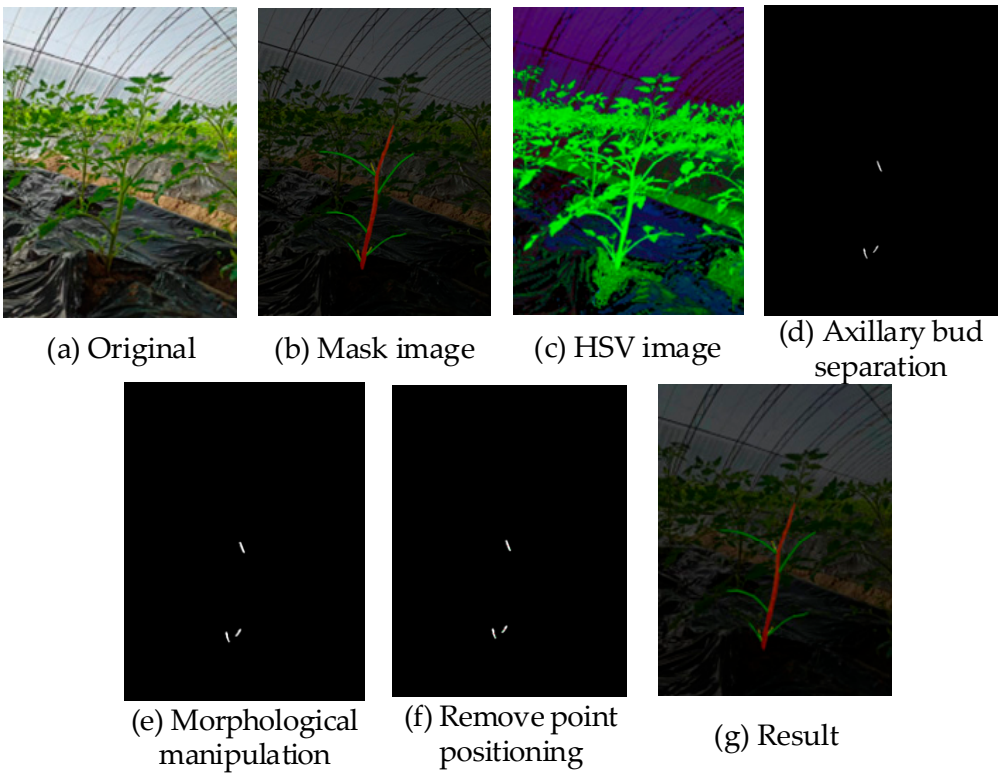


Figure 8. Example of axillary bud removal point localization.

5. Results and Analysis

This section evaluates the performance of the tomato plant multi-object recognition and positioning system based on the VGG16-UNet model. The focus is on two main aspects: segmentation performance and localization accuracy. For segmentation performance, the ability of the model to identify main stems, lateral branches, axillary buds, and background in the images will be evaluated using the intersection over union (*IoU*), mean intersection over union (*MIoU*), pixel accuracy (*PA*), and mean pixel accuracy (*mPA*) metrics. In terms of localization accuracy, two user-defined scoring criteria are used to determine whether the predicted axillary bud removal points meet agronomic requirements. This evaluation is critical to ensure that robots can accurately perform pruning tasks, which directly impacts the efficiency and quality of agricultural production.

5.1. Experimental Environment Configuration

The hardware setup used in this study is based on an Intel Core i7-8700K CPU, NVIDIA GeForce GTX 2080Ti GPU, and 64 GB of memory. The experiments were performed on

a computer running the Windows 10 Professional operating system. The development environment is based on Python 3.8, the PyCharm IDE, and the PyTorch 1.13.1 deep learning framework.

5.2. Loss Function and Evaluation Index

The loss function is used to quantify the deviation between the predicted values and the ground truth, allowing continuous model optimization towards convergence to minimize the overall loss and achieve optimal prediction results. In order to effectively balance the tomato plant and background contributions to the segmentation model loss, and thereby improve the accuracy of tomato plant segmentation, Dice Loss was used as the model training loss function, denoted by D. Dice Loss is particularly advantageous for its ability to address the class imbalance commonly encountered in semantic segmentation tasks, thereby promoting more accurate segmentation results.

$$D = 1 - \frac{2\sum_{i=1}^{N} y_i^* y_i + \varepsilon}{\sum_{i=1}^{N} y_i^* + \sum_{i=1}^{N} y_i + \varepsilon} \tag{12}$$

In the formula, N represents the total pixel value in the image, y_i^* denotes the true value of the i pixel, y_i represents the predicted value of the i pixel, and ε is a parameter adjustment value used to prevent division by zero, set to 10^{-5}.

During the training process, in addition to calculating the model loss, it is essential to select appropriate parameters to evaluate the model performance. In this study, *IoU*, *MIoU*, *PA*, and *mPA* were chosen as metrics to evaluate the segmentation performance of the model. The formulas for these specific metrics are as follows [22,23]:

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$MIoU = \frac{\sum_{i=1}^{n} (IoU_i)}{n} \tag{15}$$

$$mPA = \frac{\sum_{i=1}^{n} (PA_i)}{n} \tag{16}$$

where TP represents cases where a sample actually belongs to the positive class and the model predicts it as such. FN represents cases where a sample actually belongs to the positive class, but is misclassified as negative by the model. FP denotes samples that actually belong to the negative class, but are misclassified as positive by the model. TN denotes samples that actually belong to the negative class and are correctly classified as negative by the model. n denotes the total number of classes.

5.3. Performance Analysis of Different Semantic Segmentation Models

This study compares and analyzes five popular semantic segmentation models, namely, VGG16-UNet, Res-UNet, DeepLabv3+, PSPNet, and HRNet, to evaluate their performance in segmenting tomato plant images, as shown in Table 1. According to Table 1, UNet models with VGG16 and ResNet architectures outperform others in terms of both MIoU and mPA. This superiority indicates their effectiveness in segmenting tomato plant images. Consequently, VGG16-UNet and Res-UNet are selected as the baseline models for further research and optimization.

Table 1. Performance analysis of different semantic segmentation models.

Model	*MIoU*	*mPA*
VGG16-UNet	80.31%	88.39%
Res-UNet	75.65%	83.62%
DeepLabv3+	57.94%	72.05%
PSPNet	51.61%	54.09%
HRNet	63.48%	69.64%

5.4. Evaluation of Multi-Objective Segmentation Accuracy of Tomato Plants

In the comparative experiments, the UNet model served as the baseline, while the encoder part was replaced by two different feature extraction networks, VGG16 and ResNet, resulting in the VGG16-UNet and Res-UNet models, respectively. These models were then trained with the following settings: an initial learning rate of 0.0001, the Adam optimizer to adjust the learning rate, 200 epochs, and a batch size of 2. This setup allowed for a comprehensive comparison of the segmentation performance between the different models.

To accelerate the convergence speed of the model training and to improve its generalization ability, pretrained weights from the large ImageNet dataset were used in the training of the tomato plant segmentation model via transfer learning. As shown in Figures 9 and 10, VGG16-UNet (pretrain) represents the VGG16-UNet model using pretrained weights, while Res-UNet (pretrain) represents the Res-UNet model using pretrained weights. It should be noted that both models with pretrained weights show a significantly accelerated convergence speed during training.

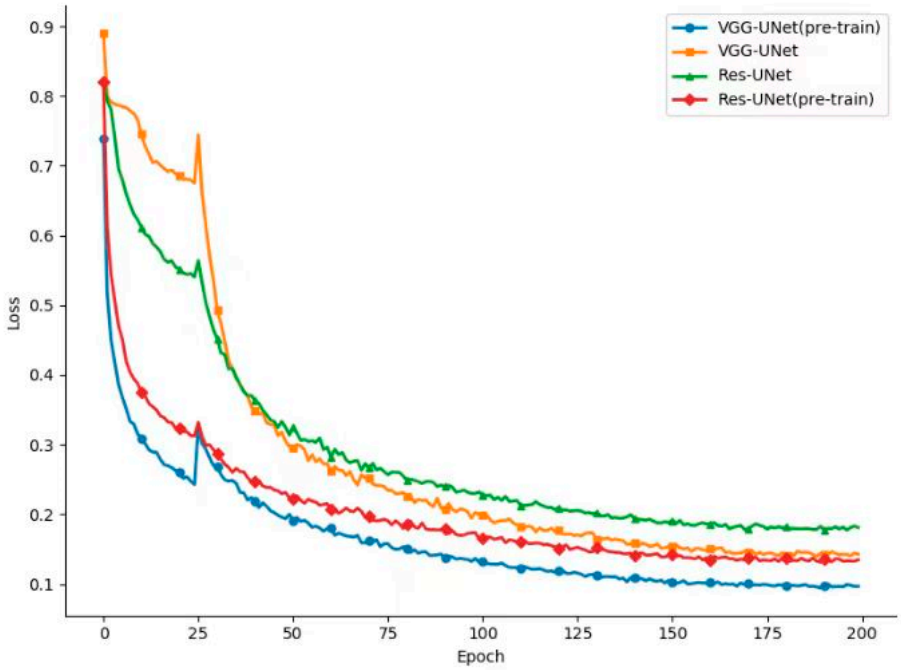


Figure 9. Comparison of convergence rate and stability during training of each model.

The analysis of Table 2 shows that the VGG16-UNet model outperforms the Res-UNet model when no pretrained weights are used. Similarly, when pretrained weights are used, the performance of the VGG16-UNet model exceeds that of the Res-UNet model. Specifically, with pretrained weights, the VGG16-UNet model shows the highest performance, achieving an average intersection over union (*IoU*) and pixel accuracy (*PA*) of 85.33% and 92.47%, respectively. This represents an improvement of 5.02% and 4.08%, respectively, over the VGG16-UNet model without pretrained weights. Furthermore, its IoU values for all segmentation categories (axillary buds, lateral branch, and main branch) are 75.91%, 78.86%,

and 87.01%, respectively, representing an increase of 10.22%, 6.73%, and 2.92% compared to the VGG16-UNet model without pretrained weights. Correspondingly, the PA values for these categories are 87.47%, 89.53%, and 93.15%, representing improvements of 8.88%, 5.55%, and 1.8%, respectively, over the VGG16-UNet model without pretrained weights.

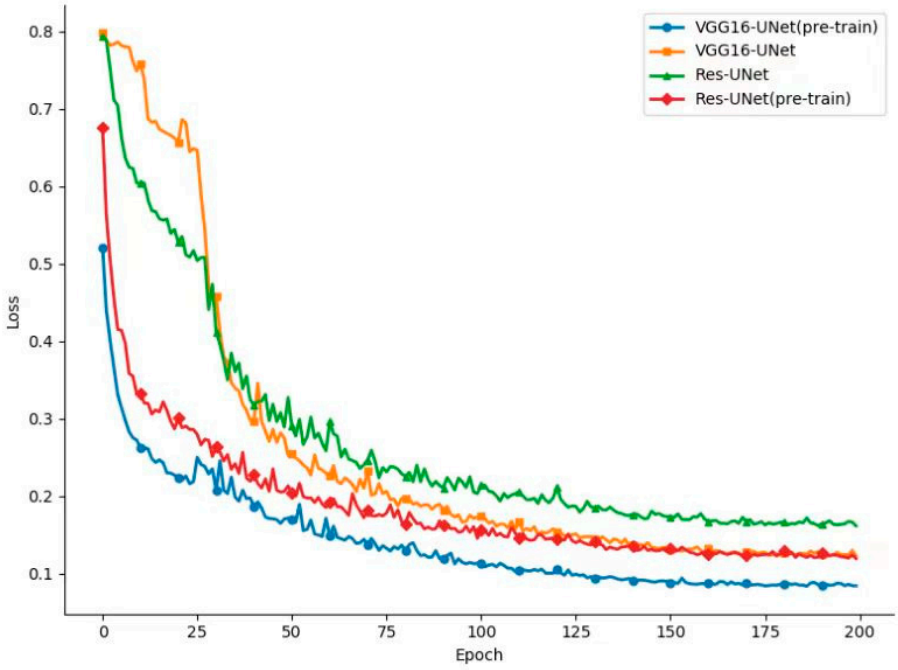


Figure 10. Comparison of convergence rate and stability during verification of each model.

Table 2. Comparison of the performance of each model.

Model	Evaluation Index	Pretrained Weights	Axillary Buds	Lateral Branch	Main Branch	Background	Mean
VGG16-UNet	IOU	False	65.69%	72.13%	84.09%	99.31%	80.31%
	PA		78.59%	83.98%	91.35%	99.65%	88.39%
VGG16-UNet	IOU	True	75.91%	78.86%	87.01%	99.54%	85.33%
	PA		87.47%	89.53%	93.15%	99.74%	92.47%
Res-UNet	IOU	False	59.58%	63.82%	80.11%	99.10%	75.65%
	PA		70.80%	76.33%	87.73%	99.61%	83.62%
Res-UNet	IOU	True	67.67%	72.23%	83.32%	99.31%	80.63%
	PA		80.09%	83.56%	90.11%	99.67%	88.36%

5.5. Comparison between Intelligent Segmentation and Manual Segmentation

To provide a more intuitive comparison of the segmentation performance of the VGG16-UNet model, the tomato plants in the test set were segmented, and the segmentation results of six images were selected for visual display, as shown in Figure 11. Overall, despite the complex and variable growth background, the VGG16-UNet model shows a high consistency between the multi-objective segmentation results of the tomato plants and the manually segmented images. Therefore, it is selected as the default model for the multi-objective segmentation and removal point localization method of tomato plants in this paper.

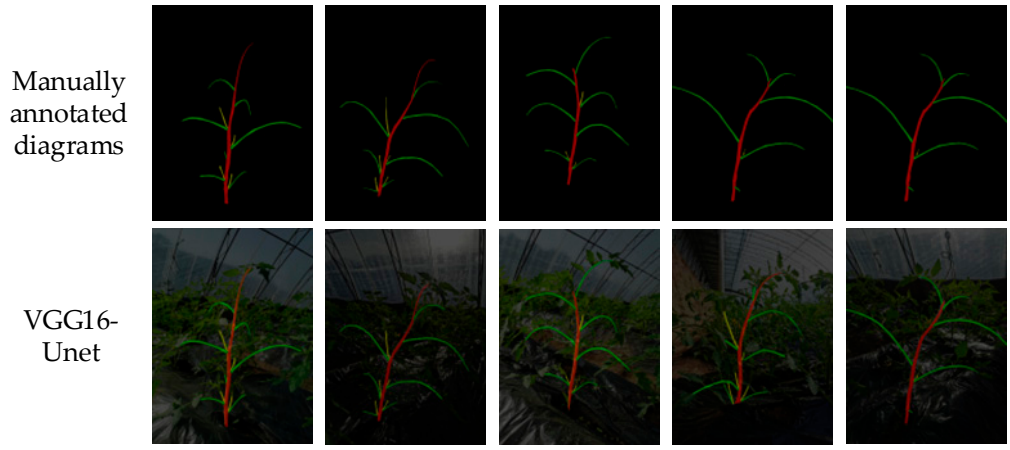


Figure 11. VGG16-UNet model segmentation visualization results.

5.6. Evaluation of the Accuracy of Axillary Bud Removal Point Positioning

When the robot end effector performs the pruning task, ensuring the accuracy of the pruning point is critical. In order to evaluate the accuracy of the pruning points, two evaluation criteria have been established as follows: Rule 1: The removal point is on the axillary branch; and Rule 2: The distance from the removal point to the junction of the main branch and the axillary branch is less than 1 cm, i.e., the stump length is less than 1 cm. Pruning points that meet both of these criteria are considered to be accurately identified.

To establish a correlation between the actual length of the axillary bud stump and the pixels in the image, the concept of pixels per inch (*PPI*) must be considered. Pixel density refers to the number of pixels per inch of screen, usually expressed in pixels per inch (*PPI*), and is calculated using the following formula:

$$PPI = \frac{\sqrt{W^2 + H^2}}{S} \tag{17}$$

where W and H are the resolution width and height of the resolution, and S is the screen size. The general formula for the actual length corresponding to the pixels in the image is as follows:

$$\text{Actual length}(CM) = \frac{2.54}{PPI} \times \text{Pixel value} \tag{18}$$

In order to match the image processing results with the practical requirements of agricultural operations, a statistical method based on pixel intervals was used, where every 50 pixels was treated as an interval. The number of pixels within these intervals and their corresponding actual lengths were statistically analyzed, as shown in Table 3. From Table 3, it can be seen that within the [0, 150] pixel interval, the actual length of the armpit is less than 1 cm, thus meeting the criteria for armpit removal. However, below the interval with the highest pixel value, we used a different partitioning method to simplify the analysis, as the pixel values in this interval no longer have any practical agronomic significance for locating armpit removal points. Calculation of the actual length corresponds to the pixel values according to the above formula. To ensure the results of the image processing correspondent with the practical requirements of agricultural operations, a statistical method based on pixel intervals was adopted, where every 50 pixels were treated as an interval. Statistical analysis was conducted on the number of pixels within these intervals and their corresponding actual lengths, as shown in Table 3. From Table 3, it can be observed that within the [0, 150] pixel interval, the actual length of the armpit is less than 1 cm, satisfying the criteria for armpit removal.

Table 3. Statistics on the correspondence between pixels and actual lengths.

Pixel Value	Actual Axillary Bud Length (cm)	Rule 2 Is Met
[0, 50]	0–0.317	True
[50, 100]	0.317–0.635	
[100, 150]	0.635–0.950	
[150–200]	0.950–1.247	False
[200–700]	1.247–4.364	

From the above calculations, the actual length in each interval can be obtained, and further evaluation can be made to determine if the agronomic requirements are met. Statistical analysis shows that within the [0, 150] pixel interval, the actual length of the axillary buds is less than 1 cm, which meets axillary bud removal Rule 2.

To assess the accuracy of the removal point calculation method, 162 images from the test set were selected for axillary bud identification. Compliance with Rule 1 and Rule 2 was recorded separately and the statistical results are presented in Figures 12 and 13 and Table 4.

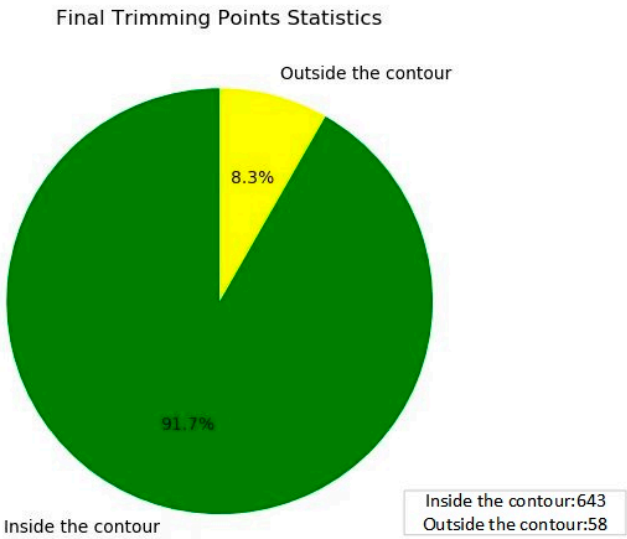


Figure 12. Rule 1 accuracy assessment.

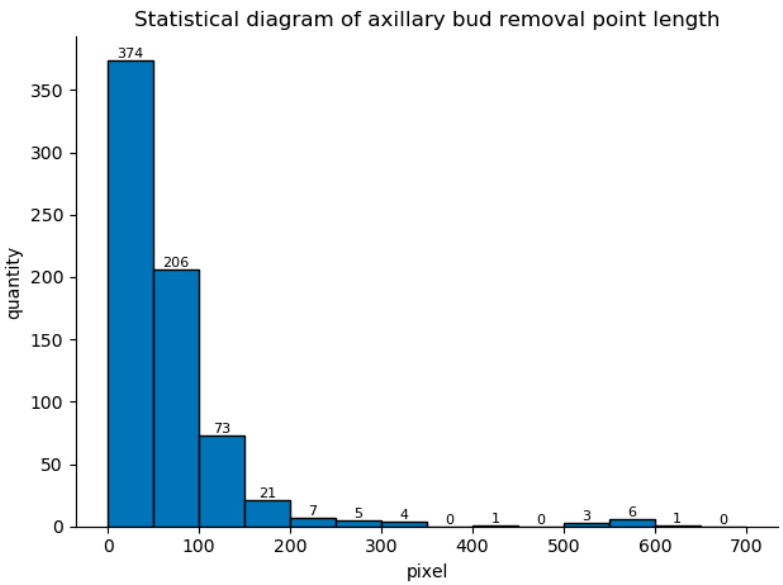


Figure 13. Rule 2 accuracy assessment.

Table 4. Axillary bud removal point identification results.

Rules	Actual Number	Eligible	Not Eligible	Accuracy (%)
Rules 1	701	643	58	91.7
Rules 2	701	653	48	93.2
Overall	-	-	-	85.5

Figure 12 uses a pie chart to statistically analyze whether all detected axillary bud removal points meet Rule 1, which refers to the number of removal points located on the axillary buds. According to the statistics in Figure 12, there are a total of 701 axillary bud objects, of which 643 removal positions are located on the axillary buds. Therefore, the recognition accuracy of Rule 1 is 91.7%.

Figure 13 shows, by means of a pie chart, whether all the detected axillary bud removal points meet Rule 2, which states that the stubble length should be less than 1 cm. Based on the statistics from Figure 13, of the 701 axillary bud objects, 653 have a stubble length of less than 1 cm. Therefore, the detection rate of Rule 2 is 93.2%.

Table 4 provides a comprehensive evaluation of the detection accuracy for Rule 1 and Rule 2. In practical applications, accuracy and compliance with agricultural standards for pruning operations by the robot end effector are only ensured when both rules are satisfied simultaneously. Based on the integrated assessment of Rule 1 and Rule 2, the overall accuracy rate for axillary bud removal points is 85.5%.

6. Conclusions

(1) A multi-target tomato plant identification model was developed. By using weights pretrained on the large ImageNet dataset to train the tomato plant segmentation model, the convergence speed of the model was accelerated. The model achieved a mean intersection over union (*MIoU*) of 85.33% and a mean pixel accuracy (*mPA*) of 92.47%, an improvement of 5.02% and 4.08%, respectively, over the VGG16-UNet model without pretrained weights. In addition, the model achieved intersection-over-union and pixel accuracies of 75.91%, 78.86%, 87.01%, and 87.47%, 89.53%, 93.15% for all the segmentation categories (axillary branch, lateral branch, and main branch), respectively. These results show an improvement of 10.22%, 6.73%, and 2.92% for *MIoU* and 8.88%, 5.55%, and 1.8% for *mPA* compared to the VGG16-UNet model without pretrained weights, allowing the accurate identification of the main, lateral, and axillary branch regions of tomato plants.

(2) Implementation of the calculation of axillary bud removal points was achieved. Building on the multi-objective segmentation of the tomato plants in VGG16-UNet, the axillary bud region of the tomatoes was identified by HSV color space conversion and the selection of color threshold ranges. Morphological dilation and erosion operations were used to remove noise and connect adjacent regions of the same target. The left and right endpoints and the centroid of the axillary buds were determined using a feature point extraction algorithm. The left and right positions of the axillary buds were then determined based on the relationship between the position of the axillary bud centroid and the position of the main branch. Finally, the axillary bud removal points were calculated using feature points based on the relationship between the axillary bud positions and the branch position.

(3) Two criteria were used to determine the accuracy of axillary bud removal points. By correlating real lengths with pixels in the images, removal points located within 150 pixels of the axillary bud pruning point, near the end of the main branch, were considered to meet the pruning length requirements. After experiments on 162 test set images, the accuracy of the axillary bud removal point localization reached 85.5%.

The next plan is to optimize the model for efficient deployment on edge computing devices. This will involve using network compression techniques or integrating advanced lightweight network architectures such as MobileNet and DenseNet to reduce the size

of the model while maintaining detection accuracy. These enhancements are critical for the practical application of technologies in areas such as intelligent agriculture, and will contribute to the automation and intelligence of agricultural production, thereby promoting the development of sustainable agriculture.

Author Contributions: Author Contributions: Supervision, J.F. and Y.Z.; conceptualization, J.F., Y.Z. and X.L.; formal analysis, X.L. and J.F.; methodology, X.L.; software, X.L.; writing—original draft, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Inner Mongolia Scientific and Technological Project under Grant (Grant No. 2023YFJM0002, 2022YFSJ0034) and funded by the Basic Research Operating Costs of Colleges and Universities directly under the Inner Mongolia Autonomous Region (Grant No. JY20220012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ongoing study.

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

References

1. Yang, M.T.; Liang, X.F. Design and kinematics analysis of a tomato branch and leaf cutting manipulator. *J. Chin. Agric. Mech.* **2021**, *42*, 8–14.
2. Li, T.; Sun, M.; Ding, X.; Li, Y.; Zhang, G.; Shi, G.; Li, W. Tomato recognition method at the ripening stage based on YOLO v4 and HSV. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2021**, *37*, 183–190.
3. Wang, H.N.; Yi, J.G.; Zhang, X.H. Research process on recognition and localization technology of tomato picking robot. *J. Chin. Agric. Mech.* **2020**, *41*, 188–196.
4. Priva. World's First Fully Automated Leaf-Cutting Robot for Tomato Crops [EB/OL]. *Priva*, 16 September 2021. Available online: https://www.priva.com/zh/blog/leaf-cutting-robot (accessed on 28 October 2022).
5. Ning, Z.; Luo, 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. (Trans. CSAE)* **2021**, *37*, 222–229.
6. Yan, Y.F. Research on the Key Technology of Branch Recognition and Location of Intelligent Chinese Wolfberry Picking Robot. Master's Thesis, Hefei University of Technology, Hefei, China, 2019.
7. Peng, H.X.; Xue, C.; Shao, Y.Y.; Chen, K.; Xiong, J.T.; Xie, Z.; Zhang, L. Semantic segmentation of litchi branches using DeepLab v3+ model. *IEEE Access* **2020**, *8*, 164546–164555. [CrossRef]
8. Peng, H.X.; Zhong, J.R.; Liu, H.A.; Li, J.; Yao, M.; Zhang, X. ResDense-focal-DeepLabV3+ enabled litchi branch semantic segmentation for robotic harvesting. *Comput. Electron. Agric.* **2023**, *206*, 107691. [CrossRef]
9. Palacios, F.; Bueno, G.; Salido, J.; Diago, M.P.; Hernández, I.; Tardaguila, J. Automated grapevine flower detection and quantification method based on computer vision and deep learning from on-the-go imaging using a mobile sensing platform under field conditions. *Comput. Electron. Agric.* **2020**, *178*, 105796. [CrossRef]
10. Afonso, M.; Fonteijn, H.; Fiorentin, F.S.; Lensink, D.; Mooij, M.; Faber, N.; Polder, G.; Wehrens, R. Tomato fruit detection and counting in greenhouses using deep learning. *Front. Plant Sci.* **2020**, *11*, 571299–571310. [CrossRef] [PubMed]
11. Wei, J.; Li, Z.; Xu, E.; Meng, Y.; Wei, H.; Wu, H. Research on hedge recognition based on DA2-YOLOv4 algorithm. *J. Chin. Agric. Mech.* **2022**, *43*, 122–130.
12. Ma, Z.Y.; Zhang, X.K.; Yang, G.Y. Research on segmentation method of rice stem impurities based on improved Mask R-CNN. *J. Chin. Agric. Mech.* **2021**, *42*, 145–150.
13. Liang, C.; Xiong, J.; Zheng, Z.; Zhong, Z.; Li, Z.; Chen, S.; Yang, Z. A visual detection method for nighttime litchi fruits and fruiting stems. *Comput. Electron. Agric.* **2020**, *169*, 105192. [CrossRef]
14. Jia, W.K.; Tian, Y.Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
15. Liang, X.F.; Zhang, X.Y.; Wang, Y.W. Recognition method for the pruning points of tomato lateral branches using improved Mask R-CNN. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2022**, *38*, 112–121. [CrossRef]
16. Liao, J.; Chen, M.H.; Zhang, K.; Zou, Y.; Zhang, S.; Zhu, D.Q. Segmentation of crop plant seedlings based on regional semantic and edge lnformation fusion. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 171–181.
17. Zhang, H.M.; Zhang, G.L.; Zhu, S.N.; Chen, H.; Liang, H.; Sun, Z.T. Remote sensing recognition method of grape planting regions based on U-Net. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 173–182.
18. Yuan, C.X.; Zhao, C.J.; Ren, Y.M.; Liu, Y.; Li, S.; Li, S. Recognition method of high-standard farmland road based on U-Net. *Trans. Chin. Soc. Agric. Mach.* **2023**, *54*, 163–169, 218.

19. Chen, M.; Jin, C.Q.; Mo, G.W.; Liu, S.; Xu, J. Online detection method of impurity rate in wheat mechanized harvesting based on improved U-Net model. *Trans. Chin. Soc. Agric. Mach.* **2023**, *54*, 73–82.
20. Guo, C.; Wang, X.; Shi, C.; Jin, H. Corn Leaf Image Segmentation Based on Improved Kmeans Algorithm. *J. North Univ. China* **2021**, *42*, 524–529.
21. Li, X.; Zhao, W.; Zhao, L. Extraction algorithm of the center line of maize row in case of plants lacking. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2021**, *37*, 203–210.
22. Zhu, Y.P.; Wu, H.R.; Guo, W.; Wu, X.Y. Identification Method of Kale Leaf Ball Based on Improved UperNet. *Smart Agric.* 2024; *epub ahead of print*.
23. Ji, Y.; Fang, J.D.; Zhao, Y.D. Clover Dry Matter Predictor Based on Semantic Segmentation Network and Random Forest. *Appl. Sci.* **2023**, *13*, 11742. [CrossRef]