applied sciences

MDPI

Article

Citrus Tree Crown Segmentation of Orchard Spraying Robot Based on RGB-D Image and Improved Mask R-CNN

Peichao Cong *, Jiachao Zhou *, Shanda Li, Kunfeng Lv and Hao Feng

School of Mechanical and Automotive Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China

* Correspondence: cpclzx@163.com (P.C.); zjc1228163074@163.com (J.Z.)

Abstract: Orchard spraying robots must visually obtain citrus tree crown growth information to meet the variable growth-stage-based spraying requirements. However, the complex environments and growth characteristics of fruit trees affect the accuracy of crown segmentation. Therefore, we propose a feature-map-based squeeze-and-excitation UNet++ (MSEU) region-based convolutional neural network (R-CNN) citrus tree crown segmentation method that intakes red–green–blue-depth (RGB-D) images that are pixel aligned and visual distance-adjusted to eliminate noise. Our MSEU R-CNN achieves accurate crown segmentation using squeeze-and-excitation (SE) and UNet++. To fully fuse the feature map information, the SE block correlates image features and recalibrates their channel weights, and the UNet++ semantic segmentation branch replaces the original mask structure to maximize the interconnectivity between feature layers, achieving a near-real time detection speed of 5 fps. Its bounding box (bbox) and segmentation (seg) AP50 scores are 96.6 and 96.2%, respectively, and the bbox average recall and F1-score are 73.0 and 69.4%, which are 3.4, 2.4, 4.9, and 3.5% higher than the original model, respectively. Compared with bbox instant segmentation (BoxInst) and conditional convolutional frameworks (CondInst), the MSEU R-CNN provides better seg accuracy and speed than the previous-best Mask R-CNN. These results provide the means to accurately employ autonomous spraying robots.

Keywords: citrus tree crown segmentation; variable spraying; mask region-based convolutional neural network; squeeze-and-excitation residual network; UNet++

Citation: Cong, P.; Zhou, J.; Li, S.; Lv, K.; Feng, H. Citrus Tree Crown Segmentation of Orchard Spraying Robot Based on RGB-D Image and Improved Mask R-CNN. *Appl. Sci.* **2023**, *13*, 164. https://doi.org/10.3390/app13010164

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 30 October 2022
Revised: 18 December 2022
Accepted: 20 December 2022
Published: 23 December 2022

1. Introduction

Traditionally, fruit trees are sprayed manually, which normally results in low efficiency, high costs, and pollution. Therefore, it is of great significance to use autonomous mobile robots for this purpose; hence, agricultural and computer scientists have teamed to make this happen [1,2]. Notably, a robot of this type must visually obtain citrus tree crown growth information so that it can apply the variable growth-stage-based spraying needs based on crown density, volume, leaf area index, and pesticide ratio [3,4].

With the rapid development of modern sensor technology, several methods of measuring tree canopy parameters have emerged. Among these, an ultrasonic sensor is a relatively cheap and effective tool. Nevertheless, to detect the entire crown, many ultrasonic sensors must be strung together to work. Moreover, bad weather and complex backgrounds quickly reduce their efficacy [5–7]. Lidar is known to accurately calculate complex physical geometric characteristics by creating three-dimensional images from the collation of a large number of single spatial measurements. However, the amount of data required to achieve a sufficiently dense point cloud is astronomical and uneconomic [8–10]. Infrared sensors are also cheap and fast, but they cannot identify tree canopy characteristics with sufficient resolution. Moreover, the slightest light pollution significantly reduces its detection performance [11,12]. Thus, vision sensors are favored due to their modest cost and rich data acquisition capability. Hočevar et al. [13], Beyaz et al. [14], and Asaei et al. [15] applied

image analysis software to red–green–blue (RGB) fruit tree pictures to separate targeted areas using an image segmentation algorithm. Canopies, backgrounds, gaps between crowns, etc. were used to improve pesticide treatment accuracy. Unfortunately, achieving "highly accurate" segmentation is terribly difficult in practice as there are seemingly infinite opportunities for noisy data to infect imagery taken from natural environments. They showed that highly accurate segmentation was key to meeting this need. Therefore, we seek to make the next milestone improvement to tree crown image segmentation efficacy and efficiency. However, we must first review the history of people trying to meet this goal.

Machine-learning image segmentation tools have already been used to measure tree crown parameters. To enable automatic apple tree canopy shape segmentation, Hočevar et al. [13] designed a machine vision system that converts RGB images to the hue–saturation–luminance color space and used a green notch filter to segment the image. An image erosion technique was then used to refine the canopy features. However, this method is extremely sensitive to noise and is not robust enough for practical use. Liu et al. [16] used the traditional watershed method of Gaussian filtering to segment tree crown boundaries, but the algorithm has high complexity and requires a huge training dataset. Gao et al. [17] used color differences to segment a variety of tree crown images according to their color characteristics, but the background colors created too much interference. Most of these methods use RGB cameras to collect two-dimensional data. Thus, they lack the required spatial information and are restricted to estimating distances based on the pre-calibrated camera range from the target. Hence, via the lack of triangulation, large spatial measurement errors are commonplace.

More recent depth camera innovations have led to exciting new studies on a variety of computer vision tasks because they can directly obtain highly precise distance information and have the advantage of secondary development platforms. However, because this technology is relatively new, it is mostly applied for 3D modeling of objects, virtual reality, and other fields; few studies have used it for canopy analysis. Xiao et al. [18] used Microsoft's Kinect depth camera for crown parameter extraction for the first time. In 2019, Milella et al. [19] used an off-the-shelf depth camera to effectively estimate grape canopy volumes and clusters. By 2020, Kim et al. [20] had developed a pear tree detection system for an orchard by using depth camera data to effectively remove background imagery noise. These researchers have conducted good work with depth cameras and tree canopies, but their methods were restricted to crown color, shape, and texture characteristics, which, even with the best cameras, are easily affected by environmental noise. It is very notable and quite surprising to some that RGB depth (RGB-D) images contain a great deal of semantic information that transcends the capabilities offered by the aforementioned banal features. Table 1 has been given to summarize and compare the use of various types of sensors for tree crowns in the above research.

Table 1. Advantages and disadvantages of the different types of sensors used for geometrical characterization of tree crown.

Sensors	Advantages	Disadvantages
Ultrasonic sensors	• Low price and good robustness • Easy to implement and good adaptability	• Large ultrasonic beam spread angle • Limited resolution and accuracy of the measurements
Lidar sensors	• Independent of environmental conditions • Provide high resolution of tree canopy structure characteristics	• The price is too high to be widely promoted • Complex structure
Infrared sensors	• Low cost and simple structure	• Influenced by light easily
Camera sensors	• Real-time and accuracy of image processing improved by machine learning significantly • Depth camera can obtain precise distance	• Highly sensitive to the weather conditions • Calibration is required, and the accuracy is not as high as Lidar

In recent years, deep learning theory has developed rapidly with the emergence of various convolutional neural networks (CNNs) that provide new and unanticipated segmentation approaches. Compared with traditional segmentation methods, this method outputs multi-size feature maps based on candidate regions, which greatly improves the ability to extract object features. In 2021, Anagnostis et al. [21] proposed an approach for orchard trees segmentation based on a deep learning CNNs, namely the Convolutional Networks for Biomedical Image Segmentation (UNet). Jose et al. [22] integrated five advanced algorithms to achieve semantic tree-crown region segmentation in complex urban scenes. Shortly afterwards, Seol et al. [23] proposed a semantic pixel-wise segmentation network (SegNet) intelligent pear-tree spraying system, which achieved the highest accuracy at the time: 83.79%. Their method is quite robust and distinguishes tree species. It can even separate the crowns from skies, buildings, and brush. Nevertheless, the continuity mask generated by the same kind of tree crown makes it impossible to distinguish different trees of the same type. Therefore, more advanced algorithms are needed; to this end, instance segmentation algorithms have been applied. Because instance segmentation can detect and segment object simultaneously, it can solve the drawback of semantic segmentation in not being able to distinguish between different individuals. In recent years, researchers have combined the instance segmentation algorithms with the use of depth cameras. For example, Liu et al. [24] developed a new instance segmentation model called "tiny Mask R-CNN", which was used to detect guava fruit and tree branches. Each detected fruit and tree branch was converted into a 3D point cloud by using the RGB camera, thus enabling the detection and 3D modeling of the fruit trees. Due to the large data obtained from the point cloud, a small number of images were collected for training. Recently, Xu et al. [25] combined depth information and improved mask-type region-based (R)-CNN (Mask R-CNN) to recognize cherry tomatoes and achieved an accuracy of 93.76% for fruit recognition, which is 11.53% higher than that obtained using standard Mask R-CNN. The experimental results of these studies show that the effective fusion of RGB-D helps to improve the feature expression ability of the model. The Mask R-CNN is a relatively advanced instance segmentation model that has been applied in the field of intelligent agriculture wildly, such as fruit detection [26], pest detection [27], evaluating ecological patterns [28], and more. Most relevant to this article is its successful application to challenging forestry tasks. For example, Safonova et al. [29] used the Mask R-CNN to recognize and segment olive tree crowns and estimate a single tree's biological volume. Accuracy of the model ranged from 77 to 95%. Hao et al. [30] achieved the simultaneous detection of Chinese fir crowns and tree heights using the method, but the precision of the detection was around 85%. Zhang et al. [31] proposed a coniferous crown segmentation and recognition method based on their improved Mask R-CNN version. However, the accuracy was improved by less than 1%. With their boundary segmentation algorithm, several geometric parameters (e.g., contour, center of gravity, and area) were amazingly extracted. Comparisons were later made with highly specialized task-specific models, such as UNet [32], which had made great advances in image-training efficiency for medical decision support, and You Only Look Once version three (YOLOv3) [33], which had a huge breakthrough in object detection. For instance, Wang et al. [34] proposed a modified YOLOv3 model to detect potholes accurately on the pavement surface. Compared with the original YOLOv3 model, the proposed model was significantly improved [35]. Different versions of YOLO models were combined with 3D ground-penetrating radar (GPR) images to recognize the internal defects in asphalt pavement by Liu et al. [36]. The Mask R-CNN astonished everyone by outperforming those and other high-profile tools in some important metrics and had high precision and strong robustness. Additionally, Mask R-CNN, an instance segmentation method, has the function of segmentation and detection. Therefore, it is more suitable for segmenting individual tree crowns than semantic segmentation methods.

In summary, current tree crown segmentation methods have the following shortcomings: it is difficult for RGB camera systems to identify specific spraying objects from many candidates [13–15]; they are not robust enough to achieve good performance when complex backgrounds are present [18–20]; most of the original Mask R-CNN is used for crown segmentation; and there is still significant room for improved accuracy [29–31].

Therefore, the goal of this paper is to provide the most precise and accurate citrus crown segmentation tool available, and we accomplish this. Our key contributions are as follows:

- For improved object identification, RGB-D images are collected, and the image noise outside the effective spraying area is removed by aligning pixels and adjusting the visual distance. Compared to that of RGB images, the bbox AP50 score is improved by 0.3% for RGB-D images.
- To increase robustness in complex backgrounds, the model is trained using citrus crown images with different backgrounds at different growth stages. The model's bbox and seg AP50 indicators are averaged over 95%, indicating a good overall performance and strong generality.
- To improve the accuracy of tree crown segmentation, an improved instance segmentation method based on the Mask R-CNN framework is proposed. The UNet++ is a commonly used semantic segmentation network [37]. We employ a feature map-based SE block (a neural network that can improve the feature extraction ability) with UNet++ (MSEU) in the R-CNN. The SE block is integrated with the residual network (ResNet) [38,39] to improve the extractability of tree crown features, and the UNet++ is introduced in the mask branch (a neural network used for segmenting images) to further improve segmentation quality. Compared with those of the optimal Mask R-CNN, the bbox and seg AP50 of MSEU R-CNN were improved by 3.4% and 2.4%, respectively.

The remainder of this paper proceeds as follows. Section 2 explains our methods and materials in producing the new image dataset and the MSEU R-CNN. Section 3 provides our analytical construct, model training plan, and findings. Then, interpretations are provided. Finally, Section 4 provides the conclusion.

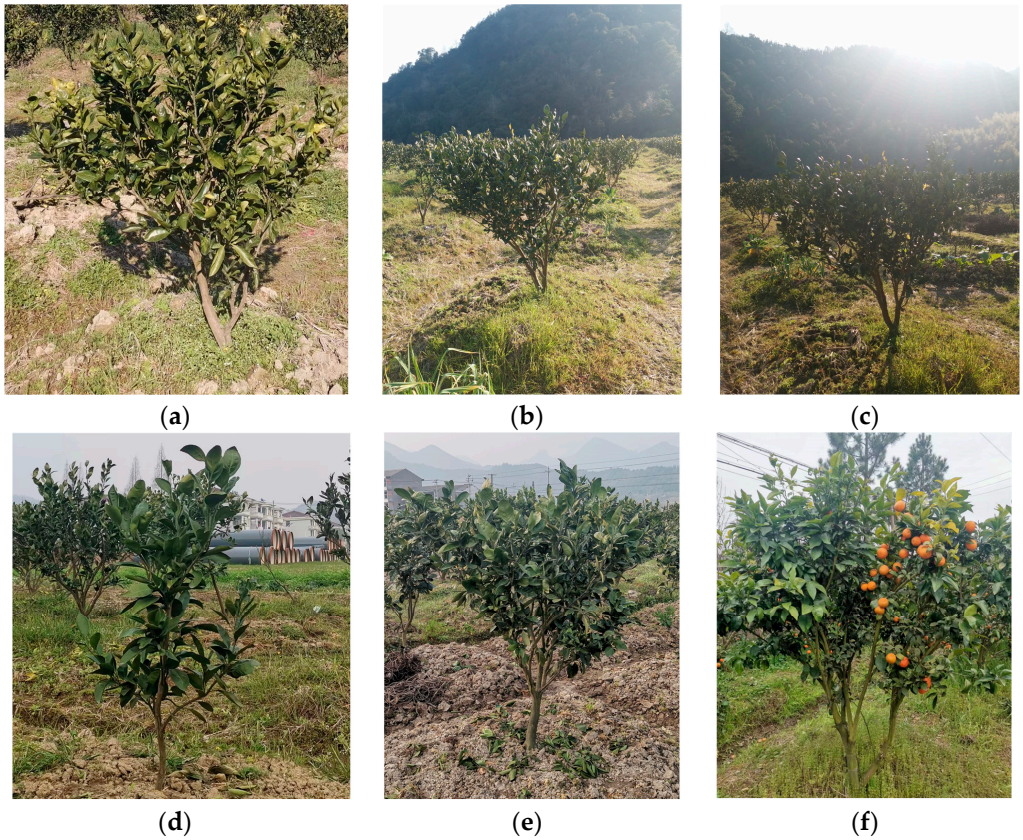
2. Methods and Materials

2.1. Image Dataset

The data in this paper were collected from a citrus plantation in the Jiaojiang District of Taizhou City in the Zhejiang Province of China. The images acquired were taken from 5–8 February 2022, from 9 a.m. to 6 p.m. A total of 766 RGB-D and RGB images of citrus trees in natural environments were collected using the RealSense d435i depth camera produced by Intel with resolutions of 1280 × 720 pixels for RGB images and 848 × 480 pixels for RGB-D images saved in the portable network graphics format. The experimental scene and image acquisition equipment are shown in Figure 1. To simulate the real working environment of a citrus spraying robot as realistically as possible, citrus tree images at different light intensities correlating with morning, noon, and afternoon periods were taken at different light angles (i.e., backlight and forward light). Various backgrounds, shooting angles (e.g., front and side views), and growth periods (e.g., seedling, flourishing, and fruiting) were collected. Example images are shown in Figure 2.

(**a**) (**b**)

Figure 1. Equipment used: (**a**) citrus plantation, (**b**) image acquisition equipment.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 2. Example citrus tree images: (**a**) forward light, (**b**) slight backlight, (**c**) severe backlight, (**d**) seedling stage, (**e**) flourishing stage, (**f**) fruiting stage.

2.2. Dataset Production

2.2.1. Generated RGB-D Tree Crown Images

The spraying robots are required to locate citrus crowns in a given planting row. To simplify the spraying process for this study, only the citrus crowns nearest the sensor on the planting line were marked as the current spraying target in each frame. According to the characteristics of orange orchard row planting, when the fruit trees in the back row are used as the background for image segmentation, they are often misjudged as spraying areas. Therefore, this paper uses the distance between the points in the tree crown depth map and the photographing center of the three-dimensional space [25] to process the depth-obtained image data. According to the distance between the camera and tree crown in the front and back scenes, the depth image segmentation method is used to eliminate the redundant image information outside the effective spraying range to block the interference

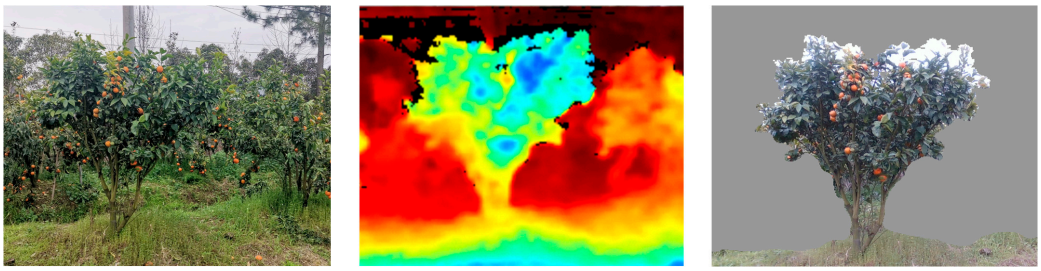
of other background trees. The RGB-D image generation process is shown in Figure 3. The RGB images (1280 × 720 pixels at 69° × 42°) and depth images (848 × 480 pixels at 87° × 58°) differ in terms of resolution and field-of-view. Therefore, in order to obtain the depth values corresponding to the pixels in the color graph, it is necessary to convert the image coordinate system of the depth image into that of the RGB image to maintain pixel-point consistency. The matrix is transformed using Equations (1)–(3) to establish a correspondence between the RGB image coordinates and depth image coordinates. Second, redundant and background information are removed by setting the depth threshold as shown in Equation (4); thus, only the foreground tree crown is retained. For example, when the camera (on the platform) is 1.2 m away from the tree, and the total depth of the tree crown (including branches and leaves) is 0.8 m, the areas of depth data exceeding 2 m are removed and shown in black.

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \tag{1}$$

$$T^{-1} = \begin{pmatrix} R^{-1} & -R^{-1}t \\ 0 & 1 \end{pmatrix} \tag{2}$$

$$T_{d2c} = T_{w2c}T_{w2d}^{-1} = \begin{pmatrix} R_{w2c}R_{w2d}^{-1} & t_{w2c} - R_{w2c}R_{w2d}^{-1}t_{w2d} \\ 0 & 1 \end{pmatrix} \tag{3}$$

$$depth(x,y) = \begin{cases} depth(x,y), & if\ d<depth(x,y) < D \\ 0, & otherwise \end{cases} \tag{4}$$



(**a**) (**b**) (**c**)

Figure 3. Tree crown RGB-D image generation process: (**a**) RGB image; (**b**) depth image; (**c**) RGB-D image.

In Equations (1)–(3), T represents the Euclidean transformation matrix; R is the rotation matrix (and also the unit orthogonal matrix); t is the translation amount of the axis; T_{w2d} represents the external parameter of the depth camera; T_{w2c} represents an external parameter of the RGB camera, and T_{d2c} represents a transfer matrix of the depth camera to the RGB camera. In Equation (4), $depth\ (x, y)$ represents the depth value of the pixel (x, y), and d and D are the two thresholds.

2.2.2. Image Annotation and Data Augmentation

In this paper, the efficient interactive segmentation (EISeg) tool [40] was used to automatically annotate the citrus crown in the RGB and RGB-D images and generate a Common-Objects-in-Context-formatted annotation file [41]. Apart from the background items, the label categories accounted for seedling, flourishing, and fruiting crown stages. Following the image annotation, to improve model generalizability, the image data were augmented using random brightness changes, contrast enhancements, and random rotations. These methods are shown in Equations (5)–(8). Therefore, to simulate the effect of noise from everywhere in the natural environment, Gaussian noise was added to the image data, as shown in Figure 4.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 4. Data augmentation process: (**a**) Original image; (**b**) Add Gaussian noise; (**c**) Rotate; (**d**) Enhance brightness; (**e**) Enhance contract.

Brightness changes: the image contrast adjustment can directly use the following transformation formula to linearly change the image in RGB space:

$$g(i,j) = f(i,j) + b \tag{5}$$

where $f(I, j)$ represents the gray level of the original pixel; the gray level of the converted pixel is $g(I, j)$. A change in coefficient b affects the brightness of the image. When b is increased, the image becomes bright, and vice versa.

Contrast enhancements: the equation for calculating the contrast can be written in the following form [42]:

$$C = k\sqrt{\frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{n=1}^{3}\left(X_{i,j,n} - \overline{X}\right)^2} \tag{6}$$

where C represents "Contrast." $R \times c \times 3$ denotes the shape of color images. The value of k is greater than 1.

Rotation: the pixel (x_0, y_0) is rotated in the original image by angle α, as shown in Formula (7):

$$[xy1] = [x_0 y_0 1]\begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

where (x, y) is the pixel after rotation.

Gaussian noise: Gaussian noise is a type of noise whose probability density function obeys the Gaussian distribution (normal distribution). The method for adding Gaussian noise can be expressed by the following formula:

$$G(x,y) = f(x,y) + n(x,y) \tag{8}$$

where $G(x,y)$ is the image with added Gaussian noise; $f(x, y)$ is the original image and $n(x, y)$ is Gaussian additive noise.

After data augmentation, images were divided into Datasets 1 and 2, which were both compartmentalized into training, verification, and testing sets at ratios of around 7:2:1. Dataset 1 contained 2000 RGB images and their annotation data, and Dataset 2 contained of the 2000 RGB-D type. See Table 2 for quantity distributions.

Table 2. Image distribution of citrus crown after data augmentation.

	Dataset 1			Dataset 2		
Growth Period	**Training Set**	**Verification Set**	**Testing Set**	**Training Set**	**Verification Set**	**Testing Set**
Seedling	455	132	65	425	152	68
Flourishing	464	146	66	479	131	67
Fruition	481	122	69	496	107	64

2.3. MSEU R-CNN Citrus Crown Instance Segmentation Model

The Mask R-CNN instance segmentation model [43] introduced a fully convolutional network (FCN)-based R-CNN [44,45] to realize pixel-level multi-target detection and segmentation. Based on the Mask R-CNN, this paper proposes the improved MSEU R-CNN to accurately segment citrus tree crowns at variable growth cycles. The overall structure is shown in Figure 5. By integrating the SE block with ResNet and combining feature pyramid networks (FPNs) [46], a backbone is formed to extract the features of the input image and to output a large number of candidate frames. Hence, the region-of-interest (ROI) in which the target may exist is filtered through a region proposal network (RPN). The ROI is then input to the ROIAlign layer and mapped to a fixed dimension feature vector via bilinear interpolation. The mapped features are then input into three branches; the tree crown is classified, and the bounding box is regressed through the fully connected layer. The Unet++ network is then used for semantic segmentation to generate a high-precision tree crown mask.

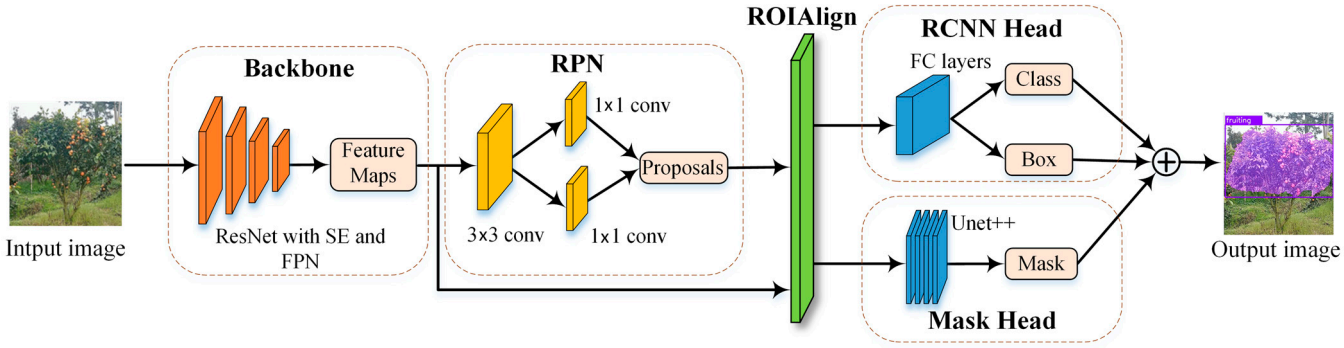


Figure 5. MSEU R-CNN network structure.

2.3.1. SE-ResNet

Figure 6a shows the ResNet adopted by the Mask R-CNN, which has good feature extractability. Nevertheless, it only focuses on the spatial information of image features and neglects the relationship between feature channels, resulting in insufficient utilization of image-feature information. The SE block provides the attention mechanism structure proposed by Hu et al. [38]. In that paper, the ResNet was optimized by embedding SE blocks to construct a new ResNet feature extraction network. Its structure is shown in Figure 6b. The SE block recalibrates the weights of different feature channels by modeling the correlations of image features and weighs them based on the previous feature channels via multiplication to enhance the attention layer to the key channel domain and suppresses ineffective feature channels. As shown in the dotted box of Figure 6b, the SE block structure mainly includes squeeze, excitation, and recalibration parts.

During the compression operation, the traditional convolution only extracts feature information in the local space, making it difficult to obtain sufficient information to characterize the relationships between channels. SE block uses a compression operation to optimize and compress the global spatial information into one channel using a global average pooling layer to achieve overall spatial feature extraction in a single channel. Mathematically, the channel is formed by compressing the feature map with the spatial dimension. Thus, the element is calculated by the following formula:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j} \tag{9}$$

where X is the input characteristic diagram; H and W are the height and width, respectively, of the characteristic diagram, and $X_{c,I,j}$ represents the elements in the channel row and column characteristic diagram matrix.

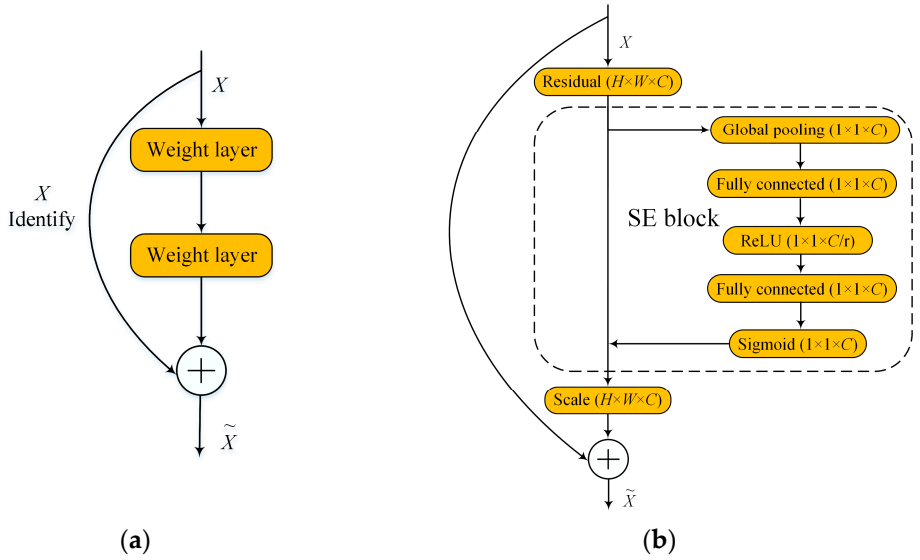


Figure 6. ResNet module and SE-ResNet module: (**a**) ResNet module; (**b**) SE-ResNet module.

In the excitation operation, the global feature map output of the compression operation is successively passed through the fully connected layer, rectified linear unit (ReLU) activation function, and Sigmoid activation function to generate the corresponding weights per channel without changing the dimensionality of the feature map. The calculation formula is as follows:

$$V = \sigma(W_2\delta(W_1 \cdot Z)) \tag{10}$$

where V is the channel weight; σ is the Sigmoid function; δ is the ReLU function; W_1 and W_2 are the weights learned from the two fully connected layers.

During the recalibration operation, the weight of the output of the excitation operation is correspondingly weighted to the previous features one-by-one through multiplication. Thus, we can realize the recalibration of the original features of each channel, thereby enhancing the attention to the key channel domain. The calculation formula of recalibration based on the input characteristic map and weight is as follows:

$$\overline{X}_C = V_c \cdot X_c \tag{11}$$

where $\overline{X}_C$ is channel feature matrix after recalibration; V_C is the corresponding weight of each channel; X_C is the channel feature matrix corresponding to each feature map.

The SE block forces the network to pay more attention to the characteristics of the citrus crown while generating a feature map. Simultaneously, via the self-attention network, the semantic information of the crown is enhanced; the complex background information of the orchard is effectively suppressed, and the problem of poor recognition accuracy when the crown is densely distributed is solved. Then the feature map output of the SE-ResNet module is used as the input to the FPN for multiscale feature extraction.

2.3.2. FPN

Because the crown of the seedling stage is much smaller than that of the flourishing and fruiting stages, to improve the detectability of small targets, the classic method uses the image pyramid method to enhance the multiscale changes of images during the training or testing stages, but it greatly increases the calculation effort of the image pyramid [46]. In this paper, the FPN method is adopted to avoid these problems, and it simultaneously better handles the multiscale change in object detection. The network structure is shown in Figure 7. The ResNet consists of five stages, convolutional layers taken from 1 to 5, respectively. Because the first layer of convolution (conv1) occupies a large part of memory, it is not included in the pyramid. Corresponding to conv2, conv3, conv4, and conv5, one

feature map with different scales is generated, expressed as C2, C3, C4, and C5 respectively. Using the feature map of the ResNet output as input to the FPN to establish the feature pyramid, the new features are output: P2, P3, P4, and P5.

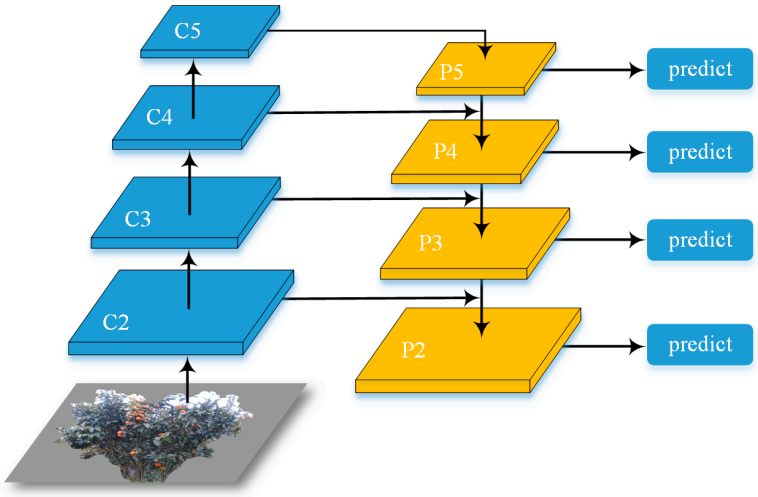


Figure 7. Feature pyramid network structure.

2.3.3. RPN and ROIAlign

The P2–P5 layer crown feature images obtained via feature extraction through the ResNet and FPN are input to the RPN, which uses 3 × 3 sliding windows on each feature map to generate candidate regions of different sizes on the original image according to three aspect ratio sets: 1:1, 1:2, and 2:1. Simultaneously, each candidate region is output as a 256-dimensional feature vector for target classification and border regression. The classification results are either tree crown or non-tree crown. After frame regression, four more accurate coordinate values are output to determine the position information of the candidate region in the original image. The candidate regions are then mapped with different input image sizes using the manually labeled crown information as the real crown region, and the intersection ratio of all candidate and real crown regions are calculated, and the candidate regions greater than the intersection ratio threshold are reserved. Finally, the candidate areas with more crown opportunities are obtained by screening the crown candidate areas through non-maximum suppression. The generated crown candidate region and its feature image are input to ROIAlign, and the corresponding ROI is clipped. ROIAlign adopts the bilinear interpolation method to adjust the above-mentioned ROIs of different sizes to a uniform size. The fixed-size ROI then passes through the fully connected layer to locate and classify the tree crown.

2.3.4. Unet++ Replaces the FCN

The mask branch of Mask R-CNN uses an FCN to extract the semantic information of the image [43], which is sensitive to the local semantic information but neglects the context information, resulting in the loss of pixel position features of the shallow network to a certain extent during image feature transmission. To better combine the shallow and deep features of the image, The MSEU R-CNN model introduces the Unet++ to improve the model segmentation performance by replacing the original mask branches. As shown in Figure 8, the Unet++ is composed of convolution blocks, down-sampling units, and up-sampling modules with skip connections between the convolution blocks. Each node in the figure represents a convolution module that combines the feature maps of four different semantic levels and makes full use of the image features of different layers to improve model generalizability. The Unet++ then redesigns the skip path based on the Unet model and uses dense skip-layer links to fuse the multiscale features of each convolution layer to achieve denser and more flexible feature propagations. From the vertical direction, each node fuses the feature images of different resolutions from the previous node to maximize

the interconnection between each feature layer. This multiscale feature fusion structure thus improves the model segmentation accuracy and convergence speed.

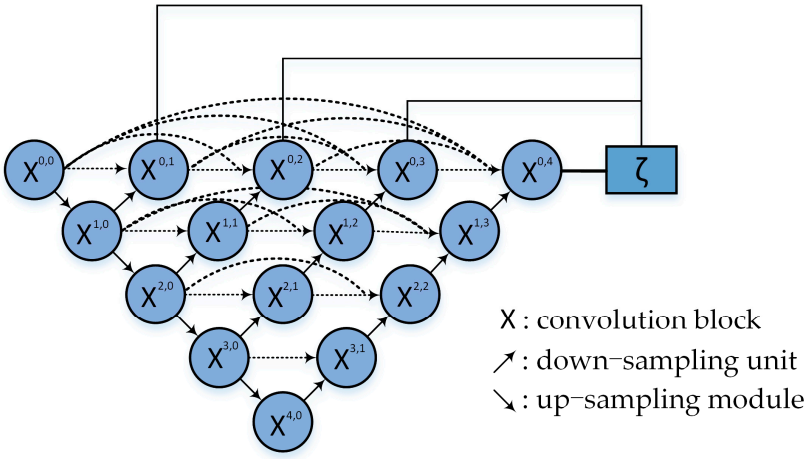


Figure 8. Network structure of Unet++.

3. Results and Discussion

3.1. Evaluation Index

To evaluate the segmentation performance of the proposed model on the citrus crown, the bbox average precision (AP50) rule and F1-score are used to evaluate the detection accuracy of the crown [47,48], in which "50" means that if the intersection over union (IoU) of the prediction and real frame are >0.5, the prediction frame is considered to be a positive sample. The bbox average recall (AR) refers to an average after [0.5, 0.95], taking the IoU of the prediction and real frame of 0.05 as the interval. The seg AP50 and the mean IoU (MioU) are used to measure the accuracy of the model for crown contour segmentation [49–51]. The seg AP refers to the average AP calculated after the IoU of the predicted and real masks are equal to 0.05 at an interval of [0.5, 0.95].

3.2. Model Training

All experiments were completed using the Windows 10 operating system. The main server configuration used a GeForce RTX2060 GPU (6 GB) and an Intel i7 9750h CPU. From this, we built a deep-learning framework using PyTorch 1.7.1 [52] and the Python 3.6.5 programming language to build the training and testing network. To speed-up model training, a CUDA graphics card v.10.1 with the cuDNN7.5 deep neural network library were installed.

The hyper parametric model training setting based on the number of photos (batch_size) was set to one; the epoch was set to 10; the stochastic gradient descent optimizer was used; the momentum factor was 0.9; the weight attenuation coefficient was 0.0001, and the initial optimizer learning rate was 0.004.

3.3. Test Result Analysis

3.3.1. RGB-D Image Validity Analysis

To understand the influence of the tree-crown RGB-D image on model detection and segmentation performance, the MSEU R-CNN algorithm was trained and tested using Dataset 1 (RGB) and Dataset 2 (RGB-D). The experimental results are shown in Table 3 where the Dataset 2 bbox AP, bbox AR, seg AP, and F1-score were 0.5809, 0.6793, 0.5747, and 0.6263, respectively, which are higher than the model performance indices trained using Dataset 1. The results show that the MSEU R-CNN model trained from the RGB-D canopy image has better detection and segmentation performance than the model trained on the RGB canopy image. According to the differential analysis of the dataset, the orchard background in the natural environment was complex and diverse,

and most of the fruit trees in the RGB image overlapped, resulting in the model becoming unsuitable for distinguishing individual tree crowns. This negatively affected the detection accuracy of the model. However, the RGB-D images included the original RGB images with color, texture, and shape tree crown and depth image information. By adjusting the visual distance and related operations, only the citrus tree crown in the foreground was retained, and most of the invalid background and nontargeted tree crown types that greatly differed from the positive samples were removed. The negative sample type required by the binary classifier can be reduced to improve the model detection accuracy. Nevertheless, the seg AP only increased by 0.1%, indicating that the RGB-D image had little impact on the model segmentation performance. These experiments confirmed that the RGB-D image reduces the interference complexity of background and non-target crown information on the detection of spraying objects.

Table 3. Validity analysis of RGB-D image.

Dataset	(Bbox)AP	(Bbox)AR	(Seg)AP	F1-Score
No using RGB-D	0.5792	0.6540	0.5737	0.6143
Using RGB-D	0.5809	0.6793	0.5747	0.6263
Promotion ratio	0.3%	4.0%	0.1%	2.0%

3.3.2. MSEU R-CNN Instance Segmentation Performance Test

To understand the performance of the MSEU R-CNN in the tree-crown segmentation task in a variety of complex backgrounds, Dataset 1 was used for model training, and citrus crown images from three growth periods were selected for testing (i.e., 200 images: 65 seedlings, 66 flourishing, and 69 fruiting). The background of each period was complex and varied. The performance of the model varies greatly in different equipment tests. To evaluate the trained model accurately, the same hardware configuration used during training was applied. The main test equipment was a high-performance computer employing a GeForce RTX2060 GPU (6 GB), an Intel i7 9750h CPU, and an SSD (100G) mainly. The fruit trees at each stage had different shapes, which differently affected the prediction results of the instance segmentation model. The MSEU R-CNN model results are shown in Table 4.

Table 4. Instance segmentation performance test results of MSEU R-CNN model.

Evaluating Indicator	Seedling	Flourishing	Fruiting	Average Value
(Bbox) AP50 (%)	95.2	99.5	95.1	96.6
(Bbox) AR (%)	66.8	82.0	70.3	73.0
(Seg) AP50 (%)	95.6	99.7	93.1	96.2
F1-score (%)	62.7	80.0	65.2	69.4

It can be seen from Table 4 that the bbox and seg AP50 of the MSEU R-CNN were 99.5 and 99.7%, respectively, which were highest of the three stages. The detection and segmentation accuracies for the seedling and fruiting stages were no more than 96%. According to the differential analysis of the crown growth stage, it can be seen that owing to the large surface areas of the citrus during the flourishing period, sufficient feature information was obtained during feature extraction, further stabilizing the training effects and enabling faster convergence. Nevertheless, the surface areas of the seedling citrus trees were small, and the characteristic information was insufficient. The training loss was therefore difficult to converge to lower levels, and the final prediction effects were poor. Although the surface areas were largest during the fruit-bearing period, and the characteristic quantities were sufficient, the surrounding weeds and other background growth were dense, causing strong interference. Owing to the great similarity between the target object and complex background features, model training was constrained, making it difficult to distinguish between contours. However, from the final test results, the average bbox AP50 of MSEU R-CNN model detection was 96.6%; that of segmentation

was 96.2%, and the recall rate was 73.0%, indicating that the overall instance segmentation performance was excellent. These experiments show that the MSEU R-CNN instance segmentation model effectively resolves the detection and segmentation problems of citrus crowns evaluated at different growth stages in unstructured environments.

3.3.3. Effectiveness Analysis of Model Structure Optimization

The proposed MSEU R-CNN model integrates SE blocks in its backbone structure. To verify its improved feature extractability, ResNet-18, ResNet-50, ResNet-101, and ResNext-50 [53] methods were combined with various Mask R-CNN models with FPN backbones for comparison. The structures of different models for segmenting citrus crown are shown in Table 5. Additionally, a comparison experiment of the model before the mask–branch improvement was added to verify the segmentation improvement of the Unet++. These models were trained and verified, by using Dataset 1. The bbox and seg AP50, AR, F1, MioU50, and reasoning scores are plotted in Table 6. From this, the MSEU R-CNN citrus crown segmentation model is shown to have an AP50 of 96.6%, which is at least 3.4% better than other models. The seg's AP was 96.2%; AR was 73.0%; F1-score was 69.4%, and the MioU50 was 74.2%, all higher than the other models. Although the inference speed of the proposed model did not change much (i.e., the average running time was 0.19 s, and the average frame rate was ~5 fps), the accuracy was significantly improved. Models 5 and 6 represent the conditions before and after the mask–branch improvement. From Table 6, it can be seen that the optimized seg AP50 and MioU50 of the mask branch reached 96.2% and 74.2%, which are 2.6% and 1.3% higher than before the improvement, respectively. This shows that the introduction of the Unet++ semantic segmentation module had a significant effect in improving the segmentation accuracy. Models 1 and 5 reflect the comparison before and after the ResNet optimization. From the data of Table 6, it can be seen that the AP50, AR, and F1-score of the bbox model after feature extraction increased by 0.6, 2.3, and 2.8%, respectively, compared with those before improvement. The ResNet layers of Models 2, 3, and 4 were deeper than those of Model 5. However, because Model 5 is embedded in the SE block, the detection accuracy index was the highest, and the feature extractability was the strongest. The comparisons between models 1–5 showed that the SE block made the SE-ResNet pay more attention to learning effective features (e.g., citrus crowns during feature-map generation while readjusting the feature channel weight value and effectively suppressing the complex background features of orchards through the network self-attention). Thus, we solved the problem of poor recognition accuracy when the crown is densely distributed. Model 5 uses a more lightweight ResNet-18 network than is conducive to improving model efficiency. The detection time of a single image is 0.19 s. Therefore, the MSEU R-CNN model with the SE-ResNet-18-FPN backbone is more robust in segmenting citrus crowns by comprehensively weighing the average precision and running speeds.

Table 5. Different model structures on citrus crown segmentation.

Number	Model	Backbone	Mask Branch
Model 1	Mask R-CNN	ResNet-18-FPN	FCN
Model 2	Mask R-CNN	ResNet-50-FPN	FCN
Model 3	Mask R-CNN	ResNet-101-FPN	FCN
Model 4	Mask R-CNN	ResNext-50-FPN	FCN
Model 5	Mask R-CNN	SE-ResNet-18-FPN	FCN
Model 6	MSEU R-CNN	SE-ResNet-18-FPN	Unet++

Table 6. Evaluation results of different model on citrus crown segmentation.

Number	(Bbox)AP50 (%)	(Bbox)AR (%)	(Seg)AP50 (%)	F1-Score (%)	MioU50 (%)	Run-Time (s)
Model 1	93.2	70.6	91.3	65.6	64.6	0.17
Model 2	77.9	57.3	76.7	52.0	61.0	0.18
Model 3	80.7	57.8	75.5	54.1	56.6	0.21
Model 4	80.5	61.4	77.4	57.1	71.0	0.23
Model 5	93.8	72.9	93.6	68.4	72.9	0.19
Model 6	96.6	73.0	96.2	69.4	74.2	0.19

Figure 9 compares the segmentation results of the different models at different growth stages. The differences in model segmentation results can be seen in the blue boxes. In Figure 9, the masks generated by the crown segmentations of each stage of Models 1–5 are missing or redundant, whereas the segmentation mask, MSEU R-CNN, is closer to the label diagram, indicating that the model's segmentation effect is obviously better than Model 1–5. In this paper, the semantic segmentation Unet++ multi-feature fusion model is used to replace the original FCN structure, maximize the relationship between each feature layer, and make full use of the image features of each layer so that the complex edges of citrus trees can be well-preserved. These experiments confirmed the feasibility of further improving the quality of crown mask segmentation by using the Unet + + multi feature fusion strategy.

To better verify the superiority of the segmentation performance of the MSEU R-CNN model, the changes to various evaluation indices during the training of different models are drawn as curves, including one for bbox AP50, and seg AP50 (Figure 10). It can be seen from the change trend of various indicators that the evaluation indicators of the MSEU R-CNN model are better than others in training, showing faster convergence and stronger robustness. Because the model proposed in this paper uses the information fusion method of the SE block to enhance its efficacy, suppress invalid features, and fully fuse the feature map information, its feature extractability is further improved so that the accuracy converges satisfactorily. Moreover, precision–recall curves graphs of various Mask R-CNN models are presented in Figure 11. The area under the curve (AUC) value of the MSEU R-CNN model is 0.9051, which is the largest value among those of the six models, indicating that the classifier works best.

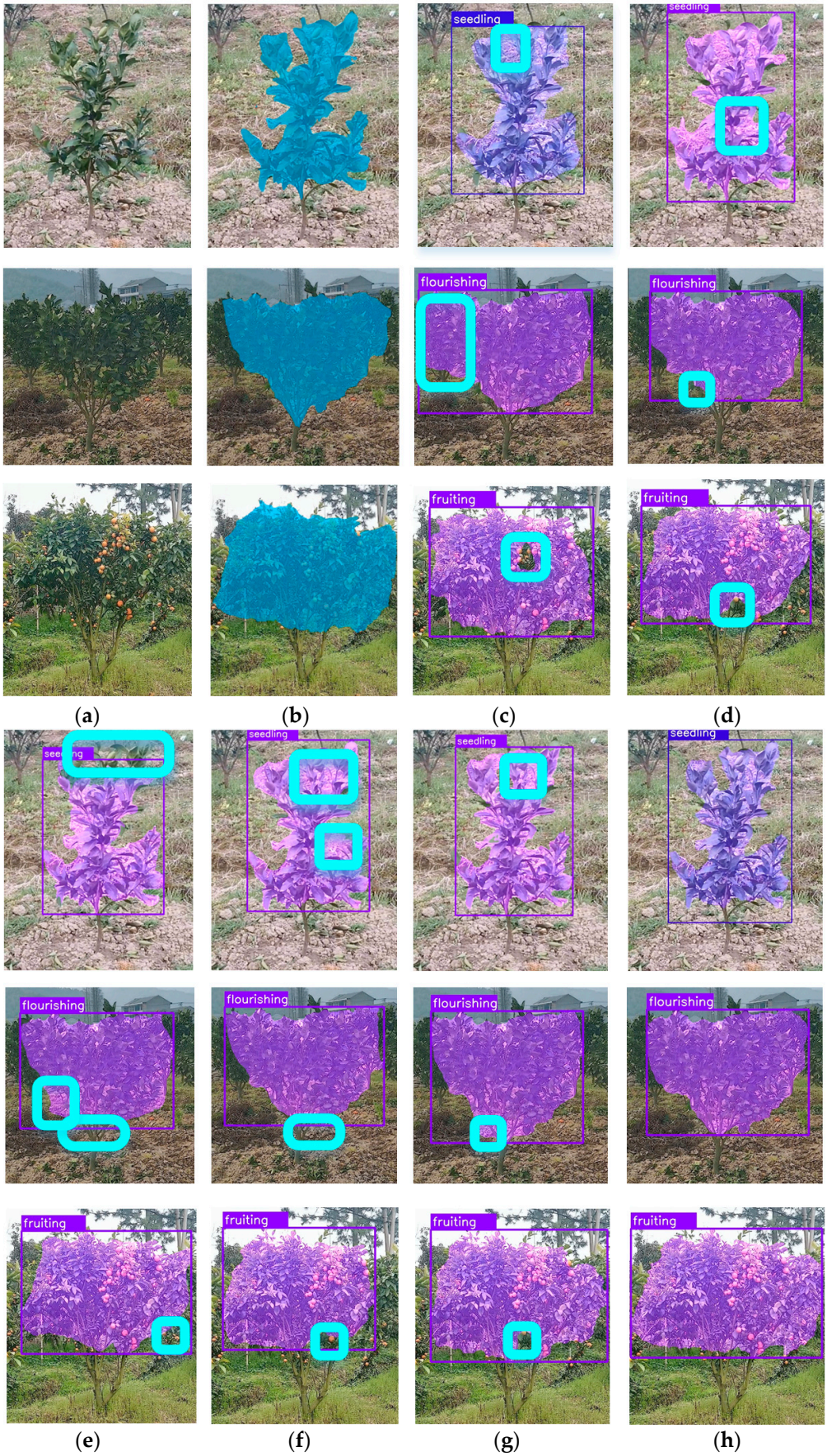
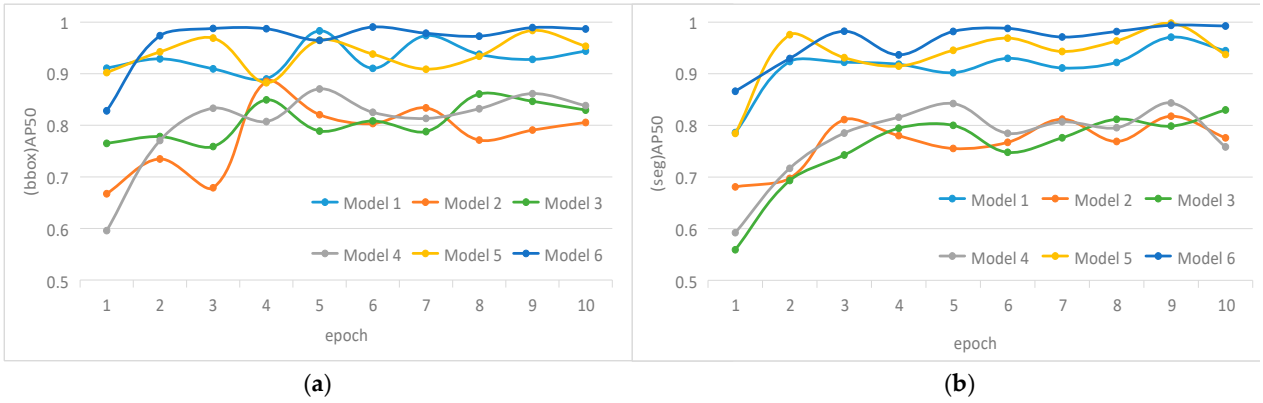
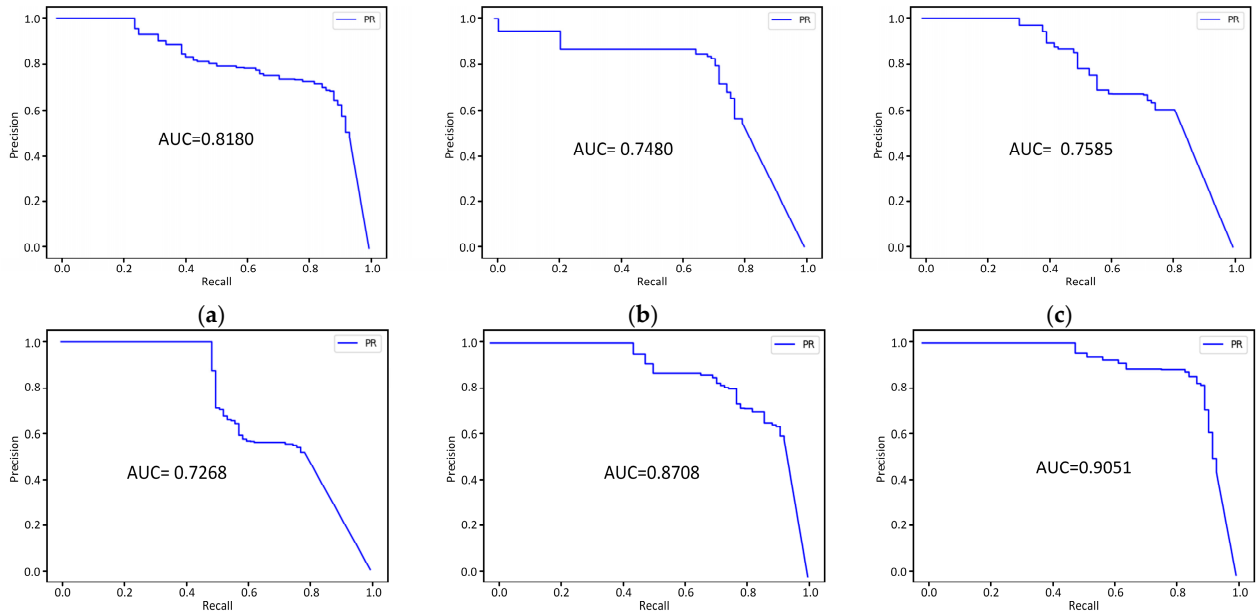


Figure 9. Segmentation effects of different models on Citrus tree-crown at different growth stages: (**a**) Original; (**b**) Label; (**c**) Model 1; (**d**) Model 2; (**e**) Model 3; (**f**) Model 4; (**g**) Model 5; (**h**) Model 6.



(**a**) (**b**)

Figure 10. Change curve of evaluation index during training stage before and after Mask R-CNN optimization: (**a**) change curve of bbox AP50; (**b**) change curve of seg AP50.



(**a**) (**b**) (**c**)

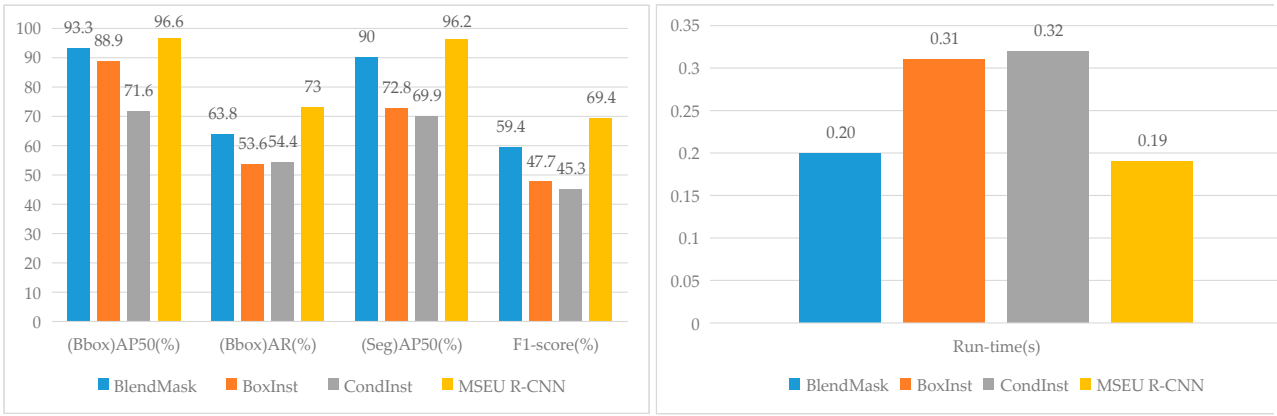
(**d**) (**e**) (**f**)

Figure 11. Precision-recall curves graphs during testing stage before and after Mask R-CNN optimization: (**a**) Model 1; (**b**) Model 2; (**c**) Model 3; (**d**) Model 4; (**e**) Model 5; (**f**) Model 6.

3.3.4. Comparison and Analysis with Other Instance Segmentation Models

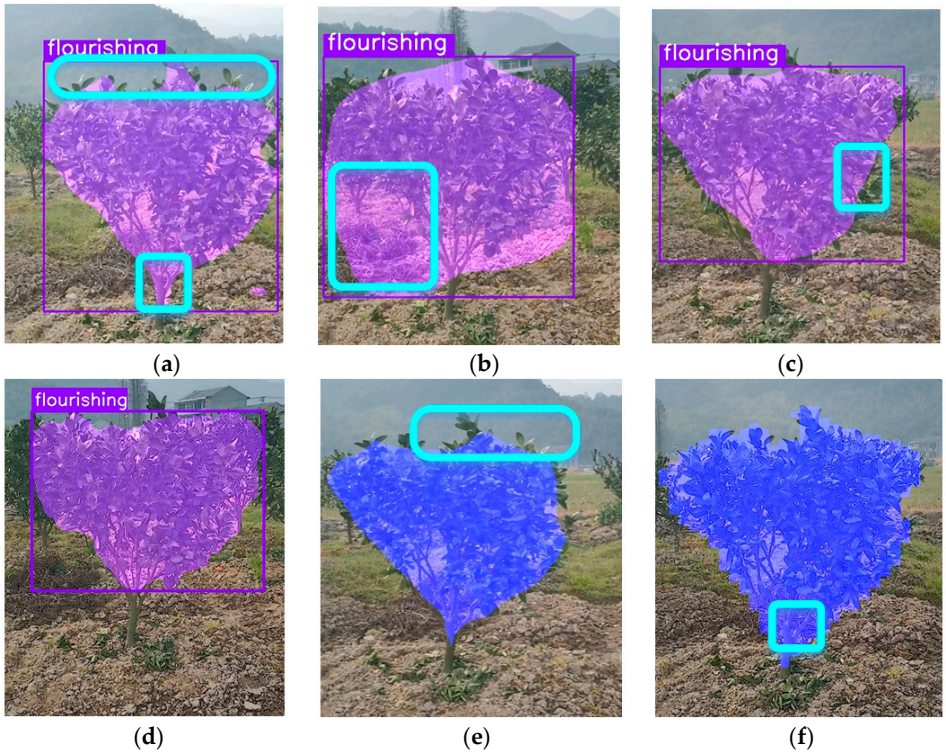
To further analyze the superiority of the MSEU R-CNN model, this paper compared it with three typical instance segmentation models: BlendMask [54], BoxInst [55], and CondInst [56]. To achieve a fair comparison, the training, verification, and testing sets of Dataset 1 were used for training and testing for all models. Figure 12 shows the performance of each model in terms of accuracy and speed. It can be seen from Figure 11 that the bbox AP50 of the MSEU R-CNN was 7.7 and 25% higher than those of BlendMask, BoxInst, and CondInst. The bbox AR was 9.2, 19.4, and 18.6%, the seg AP50 was 6.2%, 23.4%, and it was 26.3% higher. The F1-scores were 20%, 21.7%, and 24.1% higher. The results show that the proposed model had stronger feature learning ability and higher segmentation accuracy than the other three models. By analyzing the structure of MSEU R-CNN, it can be seen that the Unet++ network effectively combined the multiscale characteristics of each convolution layer and improved the segmentation accuracy to a certain extent. Compared with ResNet, the SE-ResNet paid more attention to the effective features in the image, and the feature extractability was improved. The real-time performance of the model was evaluated using the segmentation time of a single image, as shown in Figure 11. The MSEU

R-CNN only took ~0.19 s to complete an instance segmentation, and the processing time was reduced by 38.7 and 41.9%, respectively, compared with BoxInst and CondInst. Because MSEU R-CNN's residual network (SE-ResNet-18) had a shallower network layer than theirs (ResNet-50), fewer training parameters and hardware configuration were needed. Notably, the MSEU R-CNN model ran relatively fast. Moreover, Figure 13 illustrates the test results of BlendMask, BoxInst, CondInst, MSEU R-CNN, DeepLab version three (DeepLabv3) [57], and Unet on the self-defined dataset for a comparison. Because DeepLabv3 and Unet are semantic segmentation networks, they can achieve pixel-level segmentation but cannot locate the border box by instance segmentation. Further, when compared with BlendMask, BoxInst, and CondInst, MSEU R-CNN can achieve a more refined segmentation and accurate positioning. Thus, in summary, the MSEU R-CNN model proposed in this paper achieved the best model accuracy and operation efficiency compared with other models.



(**a**) (**b**)

Figure 12. Performance evaluation results of three different models: (**a**) histogram of accuracy; (**b**) histogram of real-time.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 13. Comparison of test results among different algorithms: (**a**) BlendMask; (**b**) BoxInst; (**c**) CondInst; (**d**) MSEU R-CNN; (**e**) DeepLabv3; (**f**) UNet.

4. Conclusions and Future Work

Accurate detection and segmentation of a single citrus crown during its seedling, flourishing, and fruiting stages in a complex environment are crucial for an orchard spraying robot to achieve variable spraying. Therefore, in this paper, a citrus tree crown RGB-D segmentation method and an improved Mask R-CNN were proposed. The MSEU R-CNN is a Mask R-CNN with fused SE blocks and ResNet-15 that replaces the original mask branch with the UNet++. The main conclusions are as follows:

(1) The detection accuracy of the MSEU R-CNN RGB-D tree-crown image was higher than that of RGB, indicating that the depth image can effectively reduce the interference of complex backgrounds and non-targeted tree crowns.
(2) The MSEU R-CNN's segmentation results at different stages showed that the detection and segmentation accuracies of tree crowns at the flourishing stage were the highest, whereas those at the seedling and fruit-bearing stages were lower. However, the average bbox and seg AP50 measures were more than 95%, indicating that the overall performance was excellent with strong generalizability.
(3) Compared with the original Mask R-CNN model, the proposed model effectively improves the recognition and segmentation accuracies of a single citrus crown under the condition of a small average running time change, and the segmentation quality of the crown mask is more precise, which helps accurately evaluate crown parameters. Compared with other models, the experimental results show that the segmentation performance of the proposed model is obviously better than that of BoxInst and CondInst models.

In this paper, the citrus tree crown was segmented effectively, but the variety of fruit trees was narrow. Additionally, the real-time performance of the network model should be further improved. We plan to collect images of different fruit tree varieties and expand the canopy dataset to multiple conditions to further simplify the network structure and improve real-time performance.

Author Contributions: Conceptualization, P.C. and J.Z.; methodology, P.C. and J.Z.; Writing—review and editing, P.C.; writing—original draft preparation, P.C. and J.Z.; data curation, S.L. and K.L.; validation, K.L. and H.F.; formal analysis, S.L. and H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Central Leading Local Science and Technology Development Special Fund Project, grant number ZY19183003; Guangxi Key Research and Development Project, grant number AB20058001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank all of the founders and all of the reviewers.

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

References

1. Meshram, A.T.; Vanalkar, A.V.; Kalambe, K.B.; Badar, A.M. Pesticide spraying robot for precision agriculture: A categorical literature review and future trends. *J. Field Robot.* **2022**, *39*, 153–171. [CrossRef]
2. Hejazipoor, H.; Massah, J.; Soryani, M.; Vakilian, K.A.; Chegini, G. An intelligent spraying robot based on plant bulk volume. *Comput. Electron. Agric.* **2021**, *180*, 105859. [CrossRef]
3. Manandhar, A.; Zhu, H.; Ozkan, E.; Shah, A. Techno-economic impacts of using a laser-guided variable-rate spraying system to retrofit conventional constant-rate sprayers. *Precis. Agric.* **2020**, *21*, 1156–1171. [CrossRef]
4. Chen, L.; Wallhead, M.; Reding, M.; Horst, L.; Zhu, H. Control of Insect Pests and Diseases in an Ohio Fruit Farm with a Laser-guided Intelligent Sprayer. *HortTechnology* **2020**, *30*, 168–175. [CrossRef]
5. Dou, H.; Zhai, C.; Chen, L.; Wang, X.; Zou, W. Comparison of Orchard Target-Oriented Spraying Systems Using Photoelectric or Ultrasonic Sensors. *Agriculture* **2021**, *11*, 753. [CrossRef]

6. Maghsoudi, H.; Minaei, S.; Ghobadian, B.; Masoudi, H. Ultrasonic sensing of pistachio canopy for low-volume precision spraying. *Comput. Electron. Agric.* **2015**, *112*, 149–160. [CrossRef]
7. Li, H.; Zhai, C.; Weckler, P.; Wang, N.; Yang, S.; Zhang, B. A Canopy Density Model for Planar Orchard Target Detection Based on Ultrasonic Sensors. *Sensors* **2017**, *17*, 31. [CrossRef]
8. Mahmud, M.S.; Zahid, A.; He, L.; Choi, D.; Krawczyk, G.; Zhu, H.; Heinemann, P. Development of a LiDAR-guided section-based tree canopy density measurement system for precision spray applications. *Comput. Electron. Agric.* **2021**, *182*, 106053. [CrossRef]
9. Gu, C.; Zhai, C.; Wang, X.; Wang, S. CMPC: An Innovative Lidar-Based Method to Estimate Tree Canopy Meshing-Profile Volumes for Orchard Target-Oriented Spray. *Sensors* **2021**, *21*, 4252. [CrossRef]
10. Cheraïet, A.; Naud, O.; Carra, M.; Codis, S.; Lebeau, F.; Taylor, J. An algorithm to automate the filtering and classifying of 2D LiDAR data for site-specific estimations of canopy height and width in vineyards. *Biosyst. Eng.* **2020**, *200*, 450–465. [CrossRef]
11. Hu, X.; Wang, X.; Yang, X.; Wang, D.; Zhang, P.; Xiao, Y. An infrared target intrusion detection method based on feature fusion and enhancement. *Def. Technol.* **2020**, *16*, 737–746. [CrossRef]
12. Giles, D.K.; Klassen, P.; Niederholzer, F.J.A.; Downey, D. "Smart" sprayer technology provides environmental and economic benefits in California orchards. *Calif. Agric.* **2011**, *65*, 85–89. [CrossRef]
13. Hočevar, M.; Širok, B.; Jejčič, V.; Godeša, T.; Lešnika, M.; Stajnko, D. Design and testing of an automated system for targeted spraying in orchards. *J. Plant Dis. Prot.* **2010**, *117*, 71–79. [CrossRef]
14. Beyaz, A.; Dagtekin, M. Comparison effectiveness of canopy volume measurements of citrus species via arduino based ultrasonic sensor and image analysis techniques. *Fresenius Environ. Bull.* **2017**, *26*, 6373–6382.
15. Asaei, H.; Jafari, A.; Loghavi, M. Site-specific orchard sprayer equipped with machine vision for chemical usage management. *Comput. Electron. Agric.* **2019**, *162*, 431–439. [CrossRef]
16. Liu, T.; Im, J.; Quackenbush, L.J. A novel transferable individual tree crown delineation model based on Fishing Net Dragging and boundary classification. *ISPRS J. Photogramm. Remote. Sens.* **2015**, *110*, 34–47. [CrossRef]
17. Gao, G.; Xiao, K.; Jia, Y. A spraying path planning algorithm based on colour-depth fusion segmentation in peach orchards. *Comput. Electron. Agric.* **2020**, *173*, 105412. [CrossRef]
18. Xiao, K.; Ma, Y.; Gao, G. An intelligent precision orchard pesticide spray technique based on the depth-of-field extraction algorithm. *Comput. Electron. Agric.* **2017**, *133*, 30–36. [CrossRef]
19. Milella, A.; Marani, R.; Petitti, A.; Reina, G. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Comput. Electron. Agric.* **2019**, *156*, 293–306. [CrossRef]
20. Kim, J.; Seol, J.; Lee, S.; Hong, S.-W.; Son, H.I. An Intelligent Spraying System with Deep Learning-based Semantic Segmentation of Fruit Trees in Orchards. In Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA), Paris, France, 31 May–31 August 2020; IEEE: New York, NY, USA, 2020. [CrossRef]
21. Anagnostis, A.; Tagarakis, A.; Kateris, D.; Moysiadis, V.; Sørensen, C.; Pearson, S.; Bochtis, D. Orchard Mapping with Deep Learning Semantic Segmentation. *Sensors* **2021**, *21*, 3813. [CrossRef]
22. Martins, J.; Nogueira, K.; Osco, L.; Gomes, F.; Furuya, D.; Gonçalves, W.; Sant'Ana, D.; Ramos, A.; Liesenberg, V.; dos Santos, J.; et al. Semantic Segmentation of Tree-Canopy in Urban Environment with Pixel-Wise Deep Learning. *Remote Sens.* **2021**, *13*, 3054. [CrossRef]
23. Seol, J.; Kim, J.; Son, H.I. Field evaluations of a deep learning-based intelligent spraying robot with flow control for pear orchards. *Precis. Agric.* **2022**, *23*, 712–732. [CrossRef]
24. Lin, G.; Tang, Y.; Zou, X.; Wang, C. Three-dimensional reconstruction of guava fruits and branches using instance segmentation and geometry analysis. *Comput. Electron. Agric.* **2021**, *184*, 106107. [CrossRef]
25. Xu, P.; Fang, N.; Liu, N.; Lin, F.; Yang, S.; Ning, J. Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Comput. Electron. Agric.* **2022**, *197*. [CrossRef]
26. Zhang, C.; Ding, H.; Shi, Q.; Wang, Y. Grape Cluster Real-Time Detection in Complex Natural Scenes Based on YOLOv5s Deep Learning Network. *Agriculture* **2022**, *12*, 1242. [CrossRef]
27. Craze, H.A.; Pillay, N.; Joubert, F.; Berger, D.K. Deep Learning Diagnostics of Gray Leaf Spot in Maize under Mixed Disease Field Conditions. *Plants* **2022**, *11*, 1942. [CrossRef]
28. Love, N.L.R.; Bonnet, P.; Goëau, H.; Joly, A.; Mazer, S.J. Machine Learning Undercounts Reproductive Organs on Herbarium Specimens but Accurately Derives Their Quantitative Phenological Status: A Case Study of *Streptanthus tortuosus*. *Plants* **2021**, *10*, 2471. [CrossRef]
29. Safonova, A.; Guirado, E.; Maglinets, Y.; Alcaraz-Segura, D.; Tabik, S. Olive Tree Biovolume from UAV Multi-Resolution Image Segmentation with Mask R-CNN. *Sensors* **2021**, *21*, 1617. [CrossRef]
30. Hao, Z.; Lin, L.; Post, C.J.; Mikhailova, E.A.; Li, M.; Chen, Y.; Yu, K.; Liu, J. Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (Mask R-CNN). *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 112–123. [CrossRef]
31. Zhang, C.; Zhou, J.; Wang, H.; Tan, T.; Cui, M.; Huang, Z.; Wang, P.; Zhang, L. Multi-Species Individual Tree Segmentation and Identification Based on Improved Mask R-CNN and UAV Imagery in Mixed Forests. *Remote Sens.* **2022**, *14*, 874. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022*; Springer: Berlin/Heidelberg, Germany, 2015.

33. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
34. Wang, D.; Liu, Z.; Gu, X.; Wu, W.; Chen, Y.; Wang, L. Automatic Detection of Pothole Distress in Asphalt Pavement Using Improved Convolutional Neural Networks. *Remote Sens.* **2022**, *14*, 3892. [CrossRef]
35. Liu, Z.; Gu, X.; Chen, J.; Wang, D.; Chen, Y.; Wang, L. Automatic recognition of pavement cracks from combined GPR B-scan and C-scan images using multiscale feature fusion deep neural networks. *Autom. Constr.* **2023**, *146*, 104698. [CrossRef]
36. Liu, Z.; Wu, W.; Gu, X.; Li, S.; Wang, L.; Zhang, T. Application of Combining YOLO Models and 3D GPR Images in Road Detection and Maintenance. *Remote Sens.* **2021**, *13*, 1081. [CrossRef]
37. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet plus plus: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1866. [CrossRef]
38. 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.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
40. Hao, Y.; Liu, Y.; Wu, Z.; Han, L.; Chen, Y.; Chen, G.; Chu, L.; Tang, S.; Yu, Z.; Chen, Z.; et al. EdgeFlow: Achieving Practical Interactive Segmentation with Edge-Guided Flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada; 11–17 October 2014. [CrossRef]
41. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014.
42. Liu, Z.; Gu, X.; Wu, W.; Zou, X.; Dong, Q.; Wang, L. GPR-based detection of internal cracks in asphalt pavement: A combination method of DeepAugment data and object detection. *Measurement* **2022**, *197*, 111281. [CrossRef]
43. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
44. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
45. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
46. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017.
47. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
48. Revaud, J.; Almazan, J.; Rezende, R.; De Souza, C. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
49. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
50. Wang, D.; He, D. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [CrossRef]
51. Liu, Z.; Yeoh, J.K.; Gu, X.; Dong, Q.; Chen, Y.; Wu, W.; Wang, L.; Wang, D. Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. *Autom. Constr.* **2023**, *146*, 104689. [CrossRef]
52. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
53. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017.
54. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. Blendmask: Top-down meets bottom-up for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
55. Tian, Z.; Shen, C.; Wang, X.; Chen, H. Boxinst: High-performance instance segmentation with box annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
56. Tian, Z.; Shen, C.; Chen, H. Conditional Convolutions for Instance Segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020. [CrossRef]
57. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.