agriculture

MDPI

Article

Detection and Instance Segmentation of Grape Clusters in Orchard Environments Using an Improved Mask R-CNN Model

Xiang Huang [1], Dongdong Peng [1], Hengnian Qi [1], Lei Zhou [2] and Chu Zhang [1,*]

[1] School of Information Engineering, Huzhou University, Huzhou 313000, China
[2] College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China
* Correspondence: chuzh@zjhu.edu.cn

Abstract: Accurately segmenting grape clusters and detecting grape varieties in orchards is beneficial for orchard staff to accurately understand the distribution, yield, growth information, and efficient mechanical harvesting of different grapes. However, factors, such as lighting changes, grape overlap, branch and leaf occlusion, similarity in fruit and background colors, as well as the high similarity between some different grape varieties, bring tremendous difficulties in the identification and segmentation of different varieties of grape clusters. To resolve these difficulties, this study proposed an improved Mask R-CNN model by assembling an efficient channel attention (ECA) module into the residual layer of the backbone network and a dual attention network (DANet) into the mask branch. The experimental results showed that the improved Mask R-CNN model can accurately segment clusters of eight grape varieties under various conditions. The bbox_mAP and mask_mAP on the test set were 0.905 and 0.821, respectively. The results were 1.4% and 1.5% higher than the original Mask R-CNN model, respectively. The effectiveness of the ECA module and DANet module on other instance segmentation models was explored as comparison, which provided a certain ideological reference for model improvement and optimization. The results of the improved Mask R-CNN model in this study were superior to other classic instance segmentation models. It indicated that the improved model could effectively, rapidly, and accurately segment grape clusters and detect grape varieties in orchards. This study provides technical support for orchard staff and grape-picking robots to pick grapes intelligently.

Keywords: grape; instance segmentation; Mask R-CNN; efficient channel attention; dual attention network

check for updates

Citation: Huang, X.; Peng, D.; Qi, H.; Zhou, L.; Zhang, C. Detection and Instance Segmentation of Grape Clusters in Orchard Environments Using an Improved Mask R-CNN Model. *Agriculture* **2024**, *14*, 918. https://doi.org/10.3390/agriculture14060918

Academic Editor: Signe Marie Jensen

Received: 9 May 2024
Revised: 4 June 2024
Accepted: 7 June 2024
Published: 10 June 2024

1. Introduction

As deep learning technology continues to evolve, deep learning shows great prospects in the field of agriculture, which can effectively improve agricultural productivity and promote economic growth [1]. For example, accurate detection and segmentation of fruits in complex orchard environments can realize automatic growth monitoring, yield estimation, fruit grade evaluation, and disease diagnosis of fruits. Efficient picking, cultivation, and growth monitoring of different grape varieties in complex orchard environments have a certain degree of difficulty. The similarity of some varieties of grapes leads to the difficulty in recognizing the grape varieties. Therefore, a more rapid and accurate method is explored to help the staff and grape-picking robots segment and identify the grape clusters in the orchard, upgrading the productiveness and quality of the orchard operations.

Researchers have carried out a lot of studies on fruit segmentation and classification, including grape. The traditional machine learning theory has been studied in the tasks of segmentation and classification of grapes for a long time. Maleki et al. proposed a robust algorithm based on an artificial neural network and genetic algorithm. It used color features as the basis of vineyard image segmentation. The grape clusters were segmented from foliage and background by using the near-harvested color features [2].

Liu et al. [3] used an image processing algorithm combining color and texture information to segment grapes. The proposed algorithm was combined with support vector machine to segment the grapes in the field for yield estimation. Chauhan et al. [4] used an open-source computer vision library and random forest algorithm to count, detect, and segment the blue grape clusters. However, the abovementioned grape image feature extraction methods have limitations such as generalization ability. Due to the complexity of the growing environment of grapes in the orchard, it is accompanied by natural factors such as different light intensities, fruit overlap, branch and leaf shading, etc. When there are many varieties of grape clusters in the orchard, accurate segmentation and classification of grape clusters will be greatly hindered.

As one of the symbolic algorithms of deep learning, convolutional neural network, has been successfully applied to the important aspects of fresh fruit production [5]. The convolutional neural network is the primary deep learning framework for image classification. It has many applications for the tasks of categorizing, quality control, and detection of fruits [6]. Deep learning technology has been broadly used in segmenting and classifying fruits due to its high precision and fast speed. In contrast to conventional image processing methods, deep learning technology directly takes the image as input and automatically performs feature learning and feature extraction [7]. Mohimont et al. [8] segmented white grape varieties through existing machine learning-based algorithms and encoder–decoder segmentation models such as U-Net. Santos et al. proposed a proximal sensing method combining cameras and computer vision to detect, track, and count wine grapes in vineyards. They also proposed a fruit counting method that used three-dimensional correlation to integrate and locate the detections in space, avoiding multiple counts and solving the occlusions [9]. Wang et al. proposed a new fruit segmentation method called SE-COTR to realize precise and real-time segmentation of green apples. This method effectively solved the problem of low precision and over-complexity of the segmentation model in which the fruit has the same color as the background [10]. Wang et al. [11] proposed a Transformer-based mask region-based convolution neural network (R-CNN) model for tomato detection and segmentation. The results show that this method could not only accurately detect and segment tomatoes, but also effectively identify tomato varieties and maturity stages. In the research of accurately segmenting grapes and detecting their maturity, Li et al. [12] proposed an improved Mask R-CNN algorithm. It built a grape clusters segmentation and maturity detection model by integrating different attention modules such as the squeezing and excitation attention (SE) module, etc. In summary, it can be seen that convolutional neural networks have a lot of applications in fruit recognition, segmentation, and classification tasks.

Techniques for segmentation and recognition of images mainly include semantic segmentation and instance segmentation. The goal of semantic segmentation methods is to classify each pixel in the image into a specific object class without caring about the specific instances of each object in the picture. Instance segmentation is to classify each pixel of each object while detecting and locating objects and generate different masks for different objects in the picture. In the complex vineyard environment, there are many overlapping grape fruits. Therefore, instance segmentation is needed to accurately segment different grape clusters. Mask R-CNN [13] is one of the most classical instance segmentation methods. It has been widely used in target recognition and segmentation of agricultural fruits and other related tasks. For example, Wang et al. [14] introduced the attention mechanism and deformable convolution into the residual blocks in the backbone network to enhance the ability of feature extraction. Jia et al. [15] replaced the backbone network with MobileNetV3 to speed up the model and improve the precision of the Mask R-CNN model. Yu et al. [16] introduced the Mask R-CNN network into a strawberry-picking robot to detect strawberry maturity. These studies all had achieved good results.

In the image segmentation task, the introduction of the attention module in the network can effectively enhance the network's ability to select and extract features. In recent years, some scholars have applied the relevant attention modules to the segmentation task of grapes. Chen et al. proposed an improved algorithm based on the Pyramid Scene Parsing Network, which embedded the convolutional block attention module (CBAM) attention module and the dilated convolution into the PSPNet model's network. The improved PSPNet model can segment grape bunches of different varieties quickly and accurately in a natural field environment [17]. Shen et al. proposed a new backbone network called ResNet50FPN-ED to improve the segmentation performance of Mask R-CNN. The effective channel attention module and dense upsampling convolution were introduced into ResNet50FPN-ED. It provides a reference for precise grape cultivation [18]. Therefore, when improving the Mask R-CNN network model, it is advisable to consider optimizing the network that fuses high-level and low-level features, the detection head, and the segmentation head.

The main objective of this study was to provide a fast and accurate effective method for detecting and segmenting different kinds of grapes in complex orchard environments. Based on the successful studies of previous researchers, an improved Mask R-CNN method for grape cluster instance segmentation was proposed. The main objectives to be accomplished are as follows:

(1) The ECA attention module was integrated into the ResNet-50 backbone network of Mask R-CNN to strengthen the feature selection and extraction capabilities of the network.

(2) To further improve the precision of segmentation, the dual attention network (DANet) was introduced at the head of the mask to enhance the segmentation capability.

(3) In this study, a homemade orchard grape dataset was used to validate the validity of the modified Mask R-CNN model and compare it with other classical instance segmentation models.

(4) This study explored the impact of ECA and DANet modules on other classic instance segmentation models, providing some reference for model improvement.

2. Materials and Methods

2.1. Image Dataset

In this study, the grape images were collected in a local orchard in Huzhou, Zhejiang Province, China, and were captured using a mobile phone (Redmi Note 12 Turbo, Xiaomi Communication Technology Co., Ltd., Shenzhen, Guangdong Province, China). The image acquisition time was 12:00 to 14:30. The image resolution was 3472 × 4624 pixels, and they were saved in JPG format. To improve computational effectiveness and decrease training time, the images were uniformly scaled to 1024 × 1024 pixels by the region interpolation method of the resize() function in OpenCV. The purpose of this study was to identify and segment grape clusters in a complex orchard environment. The collected grape dataset contains 8 varieties (based on the labels in the orchard). Some grape varieties were very similar, and their colors were similar to the background color. Even with manual recognition, they were difficult to accurately identify in a short period. One issue that should be addressed was that the maturity degrees of these grape varieties were different. Therefore, this study attempted to identify similar types of grapes and other grape varieties with similar colors to the growing environment.

A total of 1166 grape pictures were taken in this study. Due to the factors such as the unstructured environment of the orchard and weather, etc., these images were collected under varying light, fruit overlap, branch and leaf shading, different shooting angles and distances, etc.

2.2. Image Annotation and Preprocessing

The acquired grape images were labeled by the geometric annotation tool in the LabelMe [19] software package 3.16.7. The annotated messages were saved in JSON format in the file. The outline of grape clusters was annotated by the geometric annotation tool using a series of points. The standard used in the annotation process was to create a complete closed mask for every grape cluster in the image. The annotated instance objects were named according to the names of eight varieties of grapes, and the unmarked part was regarded as the background. The images annotated in this study included a variety of complex situations in the orchard environment, such as grape fruit overlap, trunk branches and leaves occlusion, and background color similar to grape color, etc. For different grape instance objects of an image, different IDs were added during naming to distinguish them.

Due to the imbalance in the number of images between some grape varieties, and to enrich the dataset and improve the generalization ability of the model, the original grape data were enhanced by data augmentation. The imaging library [20] function was used to process the original segmentation dataset by randomly adjusting the image contrast, Gaussian blur, random horizontal flip, and random left–right rotation within 6°. After the above processing, the number of images of each grape species showed a relatively balanced situation. The number of all images in the grape segmentation dataset after data augmentation had increased from 1166 to 3944. The enhanced dataset was divided into three parts in the ratio of 7:2:1 for training, validation, and testing sets. The specific information is shown in Table 1.

Table 1. Grape segmentation dataset information for this study.

Category	Number 1	Number 2	Training	Validation	Testing
Hongdiqiu	100	500	357	89	54
Jumeigui	160	480	345	94	41
Nantaihu	126	504	362	91	51
Xiahei	170	510	360	108	42
Xianfeng	160	480	321	98	61
Yalishanda	120	480	327	105	48
Yongyouyihao	170	510	353	107	50
Zaotian	160	480	335	97	48
Total	1166	3944	2760	789	395

Number 1 represents the number of original datasets. Number 2 represents the number of datasets after data augmentation.

2.3. Grape Cluster Instance Segmentation Based on Improved Mask R-CNN

Mask R-CNN is an advanced model for image instance segmentation. Compared to some traditional region-based convolutional neural networks, Mask R-CNN has a faster training speed and achieves end-to-end object detection and instance segmentation tasks. This end-to-end training approach helps the model learn features better and simplifies the training process. For complex scenes like orchards, it has higher accuracy in segmentation tasks and can accurately identify and segment different object instances in orchard images. It is transformed and expanded based on the target detection network Faster R-CNN [21]. First, the RoiPool layer is replaced by the RoiAlign layer. The RoiAlign layer uses the bilinear interpolation method to map the pixel values within the ROI to a fixed-size feature map thus reducing the error. Then the mask branch that splits the object mask is added to the prediction branch on the original Faster R-CNN architecture, which can realize pixel-level mask segmentation. In this study, an improved grape cluster instance segmentation method based on Mask R-CNN was presented to quickly and precisely segment different kinds of grape clusters in the orchard. Firstly, the efficient channel attention (ECA) module was incorporated into the ResNet-50 network. The purpose was to strengthen the ability of network mining orchard fruit characteristics. Then, the DANet module was introduced in the mask branch to improve the ability of fruit segmentation. The feature maps obtained through the backbone network were input into the region proposal network (RPN), and candidate regions were generated through the region proposal network. Then, the feature was extracted from each candidate region through the RoIAlign layer to align the feature with the input correctly. Finally, the grape clusters were classified, and the boundary box regression was carried out through the full connection layer. Additionally, the segmentation masks of the grape clusters will be created by a mask branch consisting of a fully convolutional network. Figure 1 shows the model architecture of the improved Mask R-CNN for grape clusters instance segmentation. Detailed information is elaborated in the following sections.

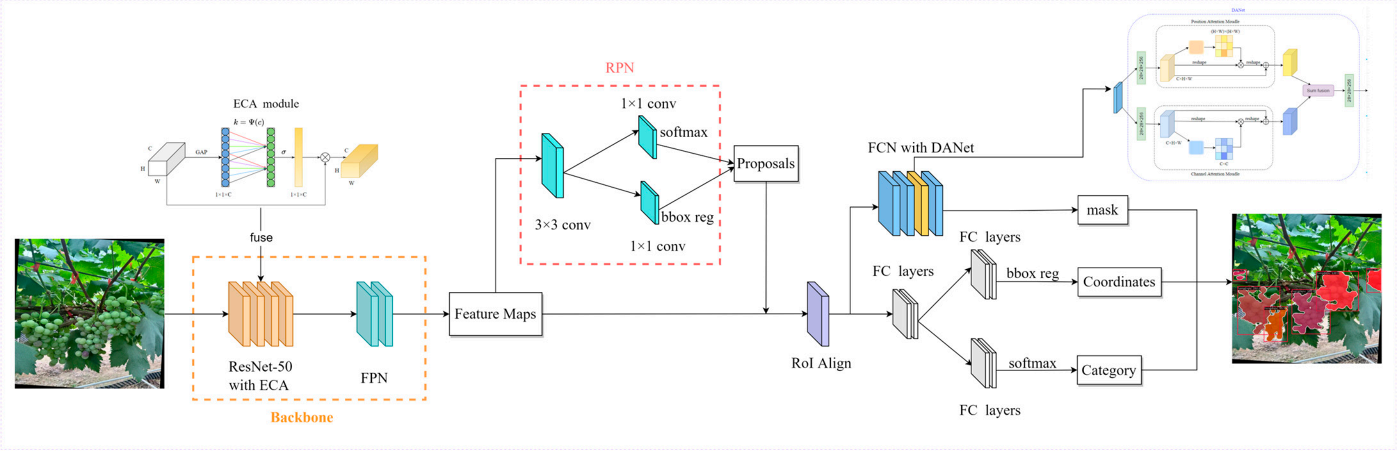

Figure 1. Improved Mask R-CNN network model architecture diagram.

2.3.1. Backbone Network

The role of the backbone network is to extract features from the input image so that they can be fed into subsequent branches of the network to carry out related tasks. In the feature mining process of the network, due to the continuous deepening of the network depth, the model performance will gradually decline. And even the problem of gradient disappearance or gradient explosion may occur. The deep residual network can effectively solve such problems. Therefore, it is introduced into the Mask R-CNN model for feature extraction.

In this study, the collected orchard grape pictures contained multiple varieties and various complex environmental factors of the orchard. Inputting grape images into the network will produce different layers of features, and the underlying features have richer detailed information, such as the outline and color of the grape pictures. High-level features have richer semantic information such as the category of grape pictures. However, high-level features lose a lot of image detail information compared to low-level features. Therefore, to better extract features of orchard grapes, the feature pyramid network [22] is used based on ResNet-50 [23] to fuse the high-level and underlying features, to better realize the feature extraction of feature maps with different scales.

In the complex environment of the orchard, the images of some of the eight grape varieties had high similarity. In addition, fruit overlapping, branch and leaf shading, and light changes in the orchard environment brought some problems to the detection and segmentation of grape clusters. Therefore, the efficient channel attention module was introduced to strengthen the extraction of effective feature information and inhibit the extraction of unimportant features. It improved the robustness of the model for identifying and segmenting grape clusters in different complex orchard environments. It was also conducive to better identification of grape varieties. The modified backbone network structure is shown in Figure 2.

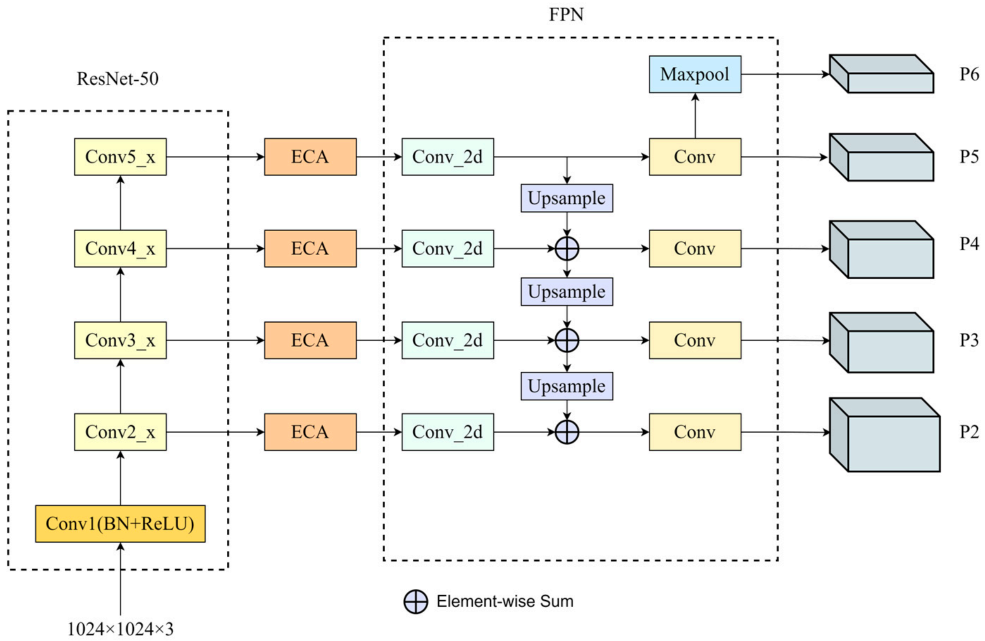

Figure 2. Improved backbone network structure diagram.

2.3.2. ECA

The ECA [24] module incorporates the strengths and compensates for the weaknesses of the SE [25] module. The ECA module utilizes one-dimensional convolution to achieve a local cross-channel interaction strategy without the need for dimensionality reduction. Its structure is shown in Figure 3. First, the input feature map is subjected to the global average pooling (GAP). Local cross-channel interactions for each channel of the feature vector and next to it are then achieved by one-dimensional convolution. At the same time, the interdependence between each channel is also obtained. The convolution kernel size k is used to determine the extent of local cross-channel interactions, which is proportional to the feature channel. It is adaptively selected by Formula (1). Among them, $|x|_{odd}$ represents the nearest odd number of x, C refers to the channel dimension, and γ and b represents relevant parameters. Finally, the channel weights after the activation function are then multiplied with the original input feature map to obtain a more refined feature map.

$$\psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

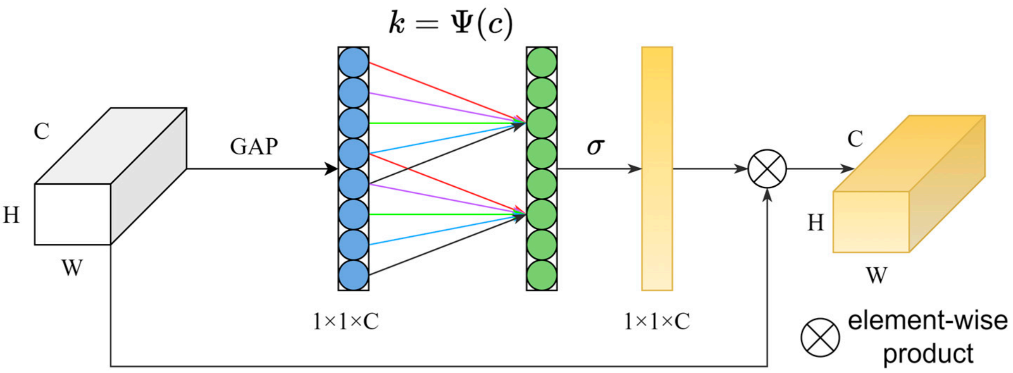

Figure 3. ECA module structure diagram.

2.3.3. DANet

To further improve the precision of grape cluster segmentation in the orchard, DANet [26] was introduced into the primitive mask head in this study. The DANet's architecture is shown in Figure 4. It is a dual attention network. By introducing the location attention module and channel attention module, the modeling ability of global semantic information and local detail information is enhanced to improve the segmentation performance. It uses these two attention modules to separately obtain the spatial relationship between feature maps at any two positions and the channel dependency relationship between channel maps. Finally, the outputs of both are summed to ascertain finer features, which is helpful to obtain more accurate instance segmentation results.

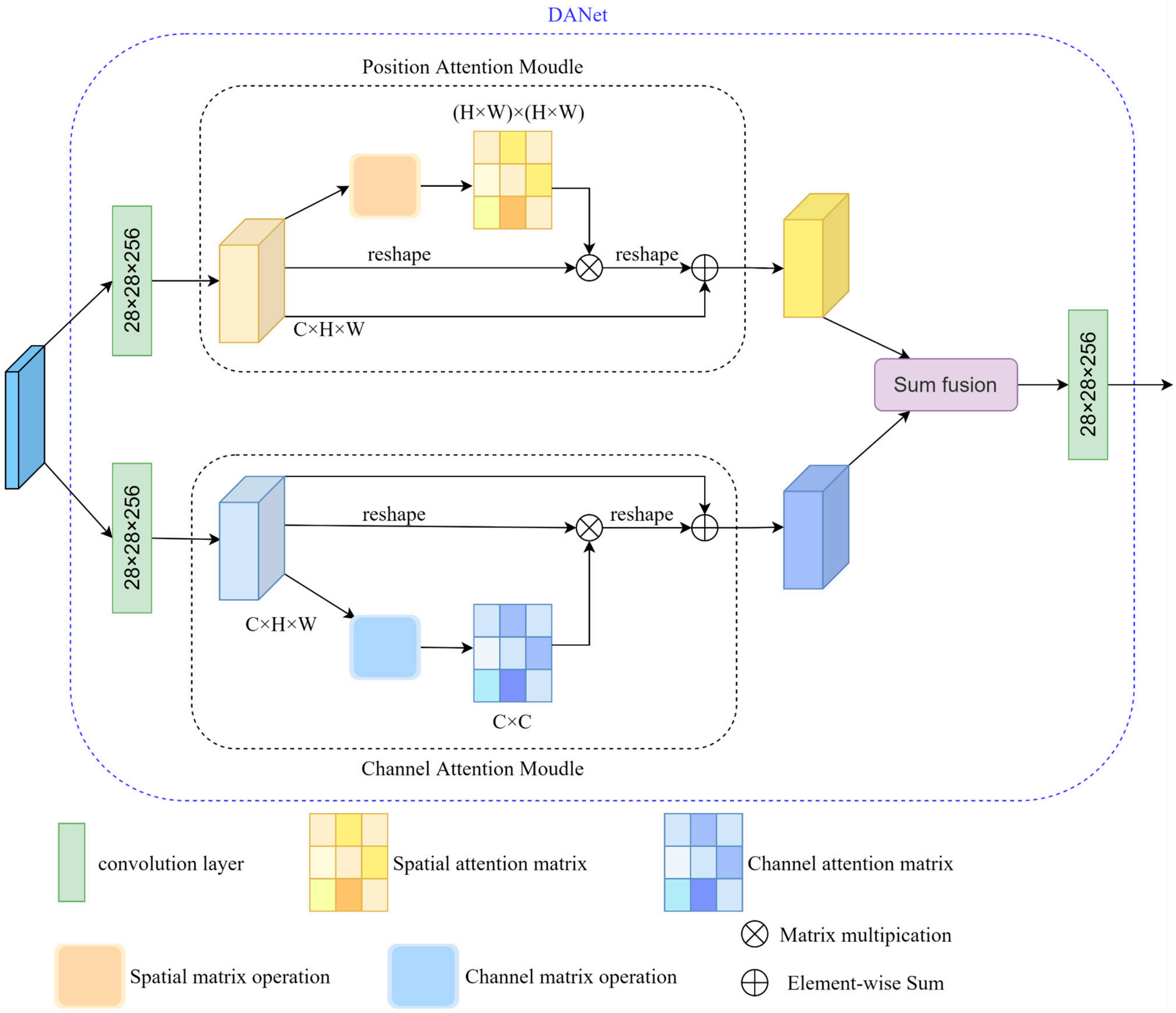

Figure 4. DANet network structure diagram.

2.4. Grape Clusters Instance Segmentation and Loss Function

The feature maps of five sizes outputted by the backbone network are inputted into the RPN network for candidate region screening. Then, the candidate areas of different sizes are mapped to the feature maps through the RoIAlign layer for alignment, so that accurate feature extraction can be performed in these areas. Next, the feature map output from the RoIAlign layer is fed into the three branch networks. The fully connected layer in the branch network is used for grape cluster localization and classification, while the fully convolutional layer is used to produce its instance mask. The loss function allows for the model to fit the data better and strengthen predictive precision. The improved Mask R-CNN loss function consists of two main parts: *RPN* loss and loss of three branches.

The branch loss includes classification, bounding box regression, and mask segmentation loss. The information is shown in Formula (2).

$$L = L_{RPN} + L_{Three-Branch} \tag{2}$$

Among them, L represents the overall loss of the improved Mask R-CNN model, and L_{RPN} represents the loss of *RPN*, whose calculation information is shown in Formula (3).

$$L_{RPN} = \frac{1}{N_{RPN\_cls}} \sum_i L_{RPN\_cls}(p_i, p_i^*) + \lambda \frac{1}{N_{RPN\_box}} \sum_i p_i^* L_{RPN\_box}(t_i, t_i^*) \tag{3}$$

Among them, L_{RPN_cls} and L_{RPN_box} refer to the classification loss and bounding box regression loss of *RPN*, respectively; λ is the weighting argument that balances these two losses. N_{RPN_cls} and N_{RPN_box} represent the quantities of samples in a small batch and the quantities of anchor positions; p_i refers to the probability of the category of anchor i; p_i^* represents the probability of the ground truth label of anchor i. When the ground truth label is positive, it represents 1. And when it is a negative label, it represents 0. t_i stands for the discrepancy between the predicted bounding box and the ground truth labeled box, t_i^* denotes the discrepancy between the ground truth labeled box and the positive label of anchor. The definition of p_i^* in the above Formula (3) is shown in Formula (4).

$$p_i^* = \begin{cases} 0, \textit{negative label} \\ 1, \textit{positive label} \end{cases} \tag{4}$$

$L_{Three-Branch}$ represents the loss of the three branches of the Mask R-CNN model. It is the summation of the losses of three task branches. The specific information of $L_{Three-Branch}$ is shown in Formula (5).

$$L_{Three-Branch} = L_{cls} + L_{box} + L_{mask} \tag{5}$$

Among them, L_{cls} is the classification loss, L_{box} is the bounding box loss, and L_{mask} is the mask loss for the segmentation mask. This research uses the same loss function as the baseline Mask R-CNN model. In addition, the classification branch will determine the category for each ROI (region of interest), the bounding box branch will determine its specific location, and the mask branch will create a fixed-size mask for each ROI.

2.5. Model Training

The whole process of this experimental study involved data acquisition, resolution reduction, dataset creation, data augmentation, model training, and model testing. The detailed experimental flowchart is shown in Figure 5. This experiment used open source platform OpenMMLab [27], Pytorch version 1.8.1, an NVIDIA RTX3090 graphics card, CUDA version 11.1, python version 3.8, and 24 GB of graphics memory. The model was trained and tested on the Windows 11 operating system. In order to promote the training and testing of the model, the labelling results of the grape segmentation dataset were converted into the labelling style of the COCO dataset. To expedite the training of the model, it was initialized by using the official weights pre-trained on the COCO dataset. Then, the data-enhanced orchard grape segmentation dataset was used to train and test the improved Mask R-CNN model. The number of training rounds and batch size were 100 and 4, respectively, the initial learning rate was 0.02, the decay weight was set to 0.0001, and the momentum was set to 0.9. In addition, the stochastic gradient descent [28] approach was used for parameter updating and training optimization.

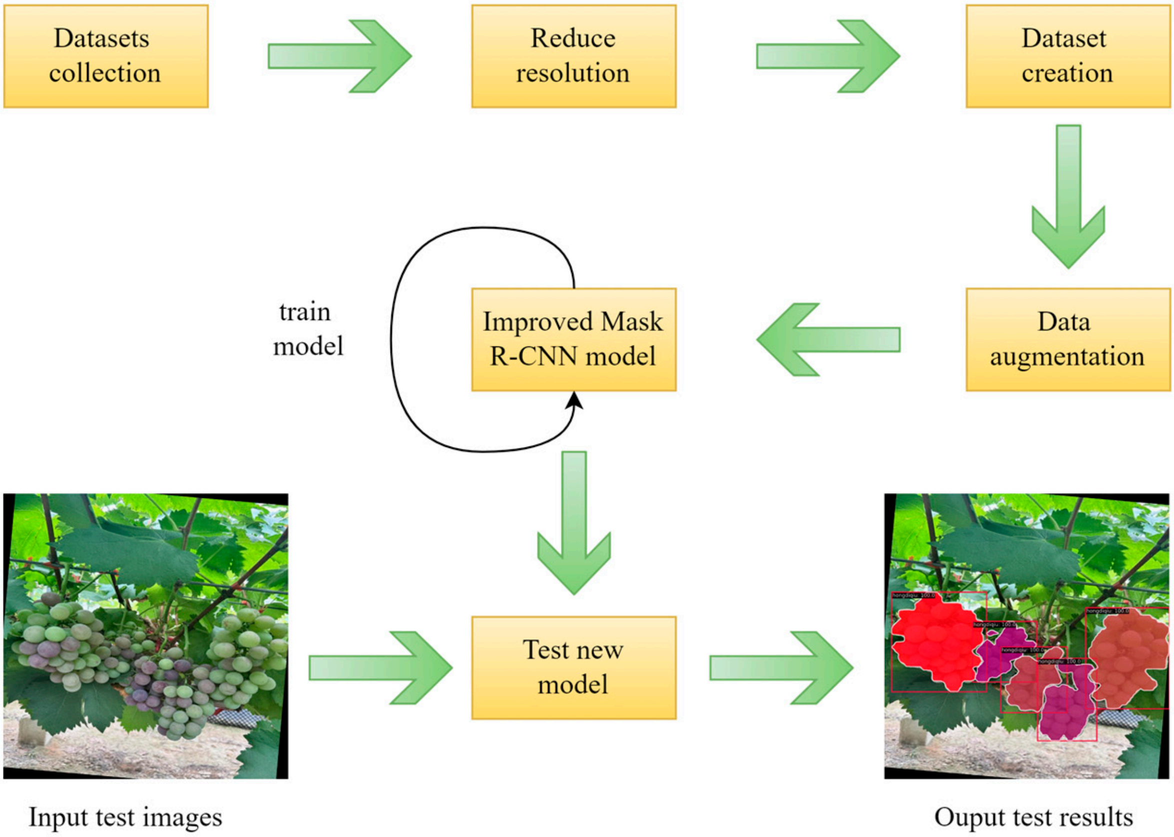

Figure 5. Overall experimental flow chart based on improved Mask R-CNN model.

2.6. Model Evaluation Metrics

Choosing the right assessment methods and evaluation metrics during the research process can help the model perform better when dealing with specific tasks. This study used COCO competitive metrics [29] involving average precision (AP) and average recall (AR) to evaluate the performance of the proposed instance segmentation model for the orchard grape cluster. The evaluation metrics also include the mean average precision for bounding box detection (bbox_mAP), the mean average precision of the segmentation mask (mask_mAP), and the average time for testing a single image. The necessary indicators in the calculation of the above evaluation indicators AP and AR include precision (P) and recall (R) are shown in Formula (6), and the calculation formula of mAP is shown in Formula (7).

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN} \tag{6}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP(i) \tag{7}$$

TP, FP, and FN in the above formula stand for true positive, false positive, and false negative samples, respectively. In instance segmentation, when the IoU of the predicted object is greater than the selected threshold, the predicted object is considered to be a true positive sample. Otherwise, it is considered a false positive sample. AP denotes the average precision of pixel segmentation and detection, and C is the total quantity of categories for instance segmentation. Some evaluation indicators of the COCO dataset used in this study are shown in Table 2.

Table 2. COCO dataset evaluation metrics.

Metric	Definition
$AP_{IoU=0.5:0.95}$	Calculate the average value of AP for every 0.05 change from IoU = 0.5 to IoU = 0.95, and use it as the final AP value
$AP_{IoU=0.5}$	Calculate AP using a threshold of IoU of 0.5
$AP_{IoU=0.75}$	Calculate AP using a threshold of IoU of 0.75
$AR_{IoU=0.5:0.95}$	Calculate the average value of AR for every 0.05 change from IoU = 0.5 to IoU = 0.95, and use it as the final AR value

3. Results

3.1. Instance Segmentation of Grape Clusters Using the Improved Mask R-CNN Model

To assess the performance of the improved Mask R-CNN model for grape cluster identification and segmentation under sophisticated background environments, the segmentation datasets of eight varieties of grapes in the orchard after data augmentation were used as the experimental datasets. The bbox_mAP and mask_mAP of the improved Mask R-CNN method on all types of grape test sets were 0.905 and 0.821, respectively. The average derivation time for a single test set image was 21.5 ms. When the IoU was 0.5 and 0.75, its (bbox_mAP, mask_mAP) could reach (0.982, 0.982) and (0.963, 0.959), respectively. The mean average recall of each category in the bounding box detection and segmentation mask were 0.933 and 0.861. The relevant results are shown in Table 3. Compared with the baseline model, the improved model has greater advantages. In the bounding box detection task, its AP value was 1.4% higher than the baseline Mask R-CNN model. When the IoU was 0.5 and 0.75, its AP value increased by 0.6% and 0.8%, respectively, compared with the baseline model. In the task of segmenting masks, the AP value of the improved Mask R-CNN model increased by 2.3% over the baseline model. For more stringent indicators, when IoU was 0.5 and 0.75, its AP value was increased by 0.8% and 2.1%, respectively, compared with the baseline model. Figure 6A,B show the comparison between the improved Mask R-CNN model and the baseline model in segm_mAP and bbox_mAP evaluation metrics, respectively. The results of the two methods showed a similar mode. The detection and segmentation precision of the first 30 epochs of the Mask R-CNN baseline model increased faster than that of the improved Mask R-CNN model. As the number of epochs increases, the detection and segmentation precision of the improved Mask R-CNN model gradually improved compared to the baseline model, and the precision of the two methods gradually stabilized after 80 epochs. It could be concluded that the improved model has a great improvement in the task of instance segmentation.

Table 3. Experimental results of the original model and improved Mask R-CNN model.

Metric	Mask R-CNN	Ours
bbox_map$_{0.5:0.95}$	0.891	0.905
bbox_map$_{0.5}$	0.976	0.982
bbox_map$_{0.75}$	0.955	0.963
bbox_AR	0.925	0.933
segm_map$_{0.5:0.95}$	0.806	0.821
segm_map$_{0.5}$	0.974	0.982
segm_map$_{0.75}$	0.938	0.959
segm_AR	0.852	0.861
Derivation time/ms	21.7	21.5
Parameters	44.01 M	44.4 M

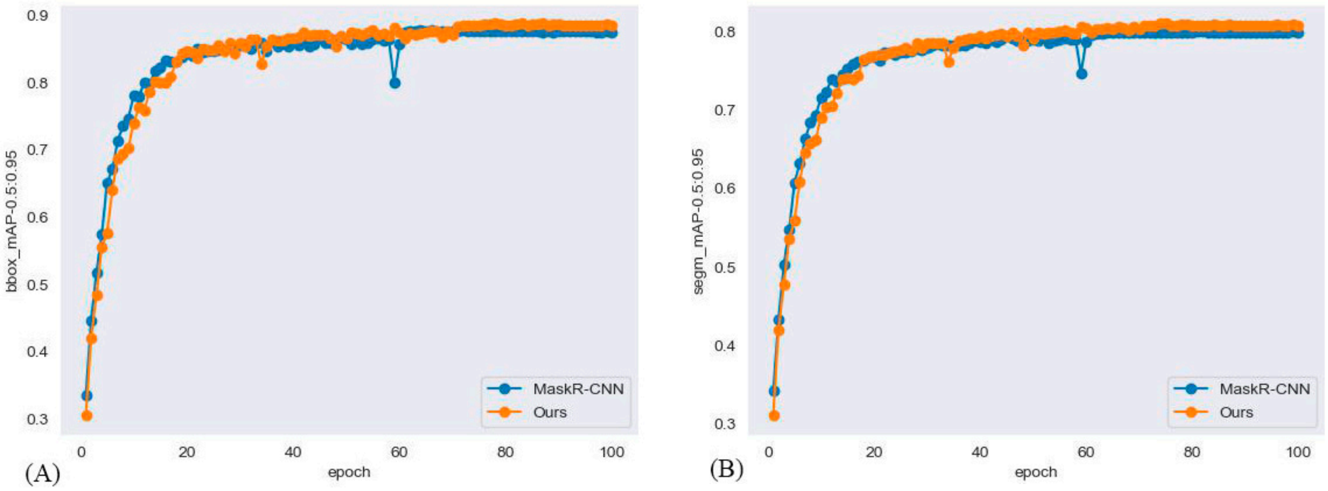

Figure 6. Comparison of the improved Mask R-CNN model and the baseline model in segm_mAP and bbox_mAP evaluation metrics: (**A**) Comparison of two methods in bbox_mAP$_{0.5:0.95}$ evaluation metric; (**B**) Comparison of two methods in segm_mAP$_{0.5:0.95}$ evaluation metric.

To further analyze the detection and segmentation results of different grape cluster varieties, the experimental results of eight grape cluster varieties are summarized in Table 4. Table 4 shows the test comparison results between the baseline Mask R-CNN model and the improved Mask R-CNN model. The category names in the first row of the Table 4 are abbreviations of the eight grape varieties. The hdg, jmg, nth, xh, xf, ylsd, yyyh, and zt refer to hongdiqiu, jumeigui, nantaihu, xiahei, xianfeng, yalishanda, yongyouyihao, and zaotian grape, respectively. From the results in Table 4, the detection and segmentation precision of these eight grape varieties can be seen. In the experimental results of the baseline Mask R-CNN model, the segmentation precision of hongdiqiu, jumeigui, and zaotian grapes will be relatively lower than other grape varieties. It also can be found that only the xiahei grape variety has a slight decrease in detection and segmentation precision in the improved model. However, the detection and segmentation precision of other grape varieties improved. The improvement was more obvious in jumeigui, xianfeng, and zaotian grapes, especially the xianfeng variety. Compared with the baseline model, the precision of boundary box detection and mask segmentation for xianfeng grape of the improved model were improved by 3.9% and 3.4%, respectively. The detection and instance segmentation visualization results of the improved Mask R-CNN for these eight grape cluster varieties are shown in Figure 7. It can be demonstrated that the improved Mask R-CNN model can accurately segment different types of grape clusters. In addition, the improved model can also achieve satisfactory detection and segmentation results for grapes in different complex orchard environments. The relevant visualization results are shown in Figure 8. The improved Mask R-CNN model can accurately detect, classify, and segment grape clusters in many cases, such as fruit overlap, branch and leaf occlusion, different light changes, fruit and background color similarity, etc.

Table 4. Comparative experimental results of detection and segmentation for eight grape varieties.

Method	Metric	hdq	jmg	nth	xh	xf	ylsd	yyyh	zt
Mask R-CNN	bbox_map	0.897	0.889	0.905	0.886	0.876	0.922	0.882	0.874
	segm_map	0.792	0.784	0.836	0.828	0.806	0.841	0.816	0.746
Ours	bbox_map	0.897	0.906	0.922	0.881	0.915	0.927	0.900	0.889
	segm_map	0.812	0.804	0.848	0.819	0.840	0.849	0.832	0.766

From the visualization results of the instance segmentation model (shown in Figure 8), the results indicated that our proposed method can accurately segment grape clusters within a variety of conditions such as uneven color (Figure 8(A1,C2,C3)), shadow influence (Figure 8(A2,A3,C2,E1,E3)), and uneven lighting (Figure 8(A2,A3,C1,C3,E2)). At the same time, this method can effectively detect and segment grape clusters that are obstructed by branches and leaves (Figure 8(C3,E1,E2)) and have overlapping fruits (Figure 8(C1,C2,E2)) accurately. It can also effectively recognize and segment the grape clusters which are divided into several parts by branches and leaves. Satisfactory detection and segmentation results can also be obtained for grape clusters with strong illumination (Figure 8(A1)) and extremely dark illumination (Figure 8(E3)).

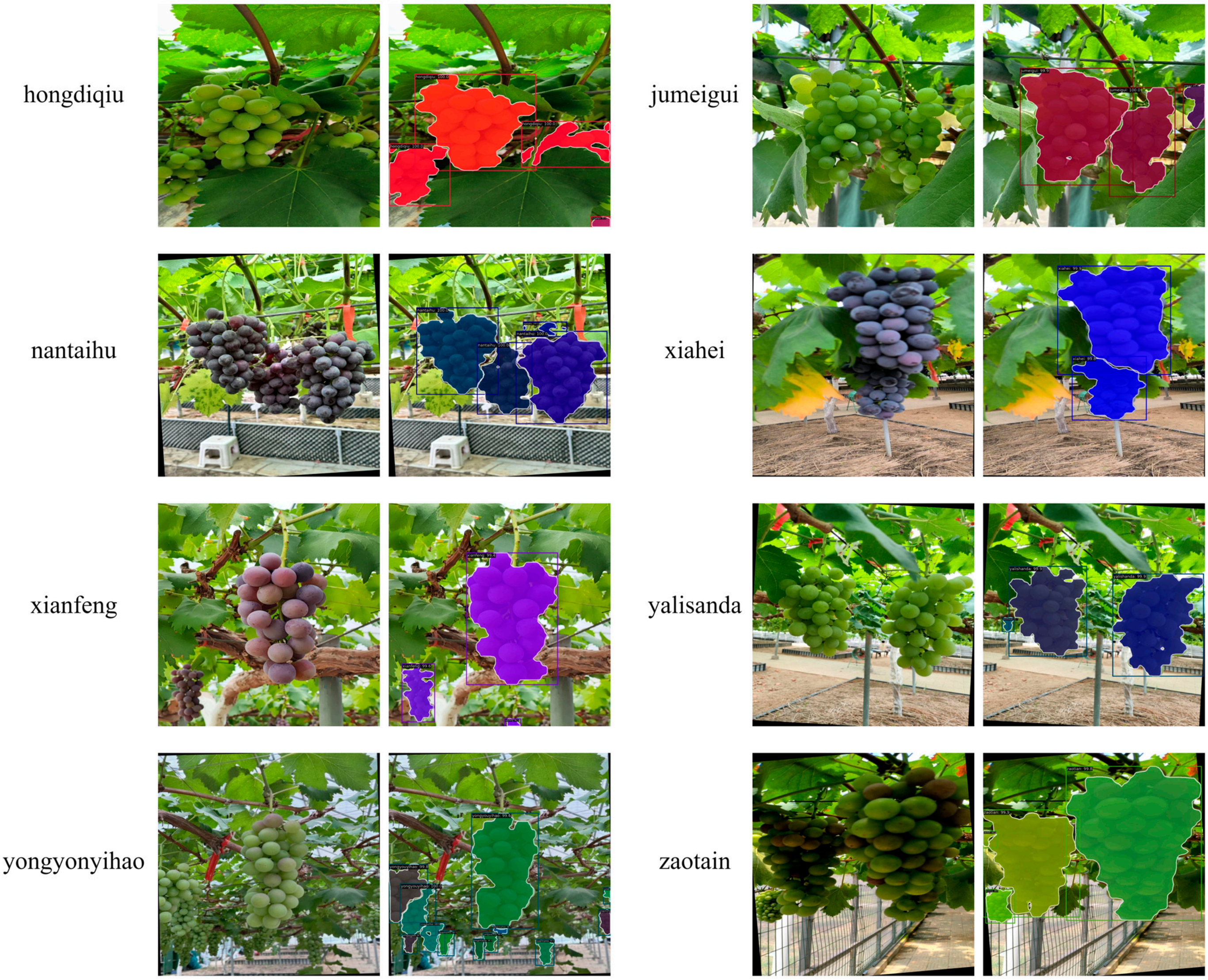

Figure 7. Visual results of detection and segmentation of eight grape cluster varieties.

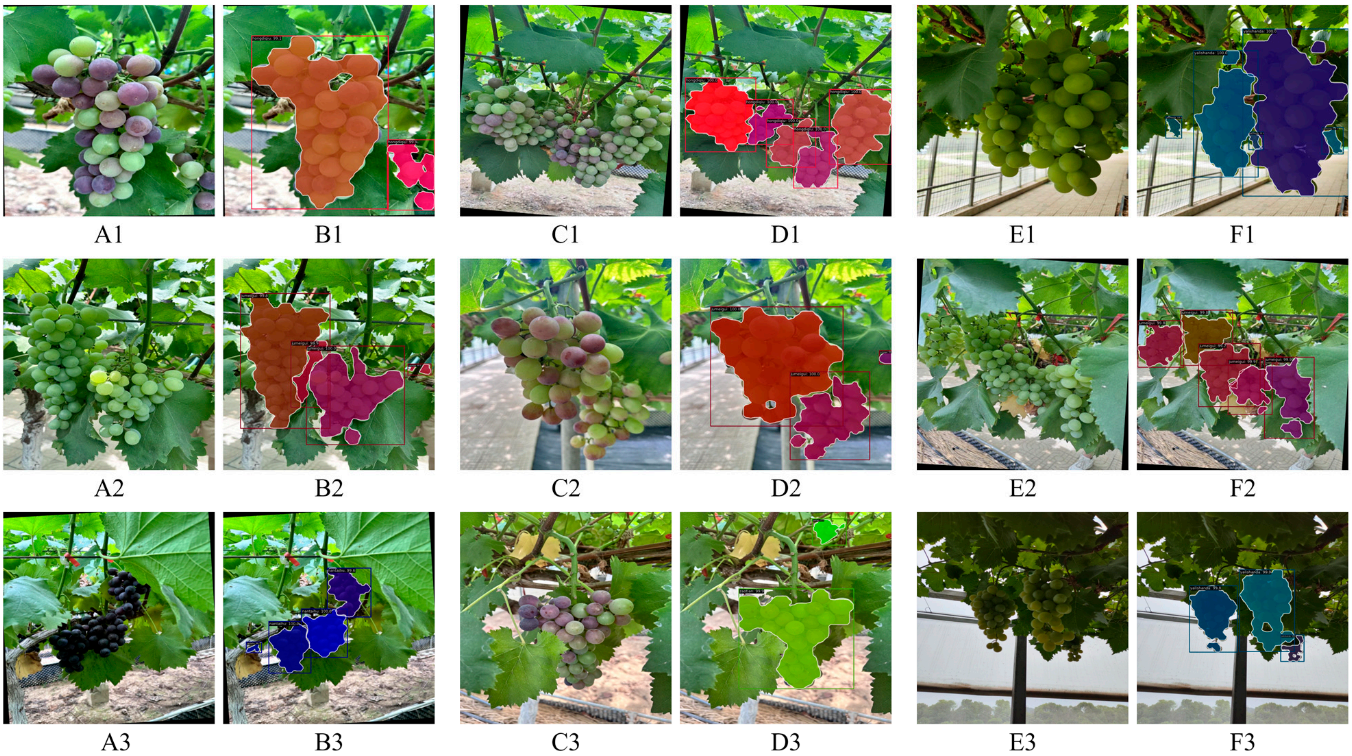

Figure 8. Examples of grape detection and instance segmentation in different complex environments: (**A**,**C**,**E**) Original images. (**B**,**D**,**F**) Visualization results of detection and segmentation of original images. (**A1**,**B1**) Grapes are affected by close-up shots, uneven color, and direct sunlight. (**A2**,**B2**) Grapes affected by uneven lighting, shadows, and similar background colors. (**A3**,**B3**) Grapes affected by uneven light and shadow. (**C1**,**D1**) Grapes are affected by density, overlapping fruit, and different light variations. (**C2**,**D2**) Grapes affected by overlapping fruit, uneven shading, and color. (**C3**,**D3**) Grapes affected by branches and leaves, uneven light, and uneven color. (**E1**,**F1**) Grapes affected by foliage and shadow. (**E2**,**F2**) Grapes affected by overlapping fruit, shading of branches and leaves, and uneven lighting. (**E3**,**F3**) Grapes affected by shadow, backlight shot.

From the comprehensive results shown in Tables 3 and 4 and Figures 7 and 8, it can be clearly illustrated that our optimized and improved Mask R-CNN model can effectively complete the segmentation and identification of eight grape varieties. It was promising to overcome the influence of various natural factors such as orchard lighting, occlusion, overlap, and shadows. It can accurately and effectively detect grape clusters with various natural factors. It also can segment and classify different grape cluster varieties accurately and has good generalization ability.

3.2. Comparison with Other Methods

To further validate the improved Mask R-CNN model, the proposed model was compared with the SOLO [30] model, Cascade Mask R-CNN [31] model, HTC [32] model, and the original Mask R-CNN model. These comparison models were also trained in the same model configuration conditions as the improved Mask R-CNN model. The comparison models were initialized with their respective official training weights on the COCO dataset. The usage configurations of these five methods are shown in Table 5. In the comparative experiment, the relevant evaluation indicators in the COCO dataset format were also used. These comparison models used the same grape segmentation dataset, divided in the same proportions as the improved Mask R-CNN model. Table 6 shows the detection and instance segmentation results of these five methods, among which the SOLO model itself did not include bounding box detection function.

As can be observed in Table 6, in comparison with the other four methods, the improved method based on Mask R-CNN can effectively improve the bbox_mAP, segm_mAP, segm_AR, and other evaluation indicators in grape cluster detection and segmentation. In addition, compared with the baseline Mask R-CNN model, the parameters of the new model are only slightly increased, but the precision of boundary box detection and mask segmentation has been greatly improved. This verified the effectiveness of our improved ECA attention module in the model, which can effectively improve the precision of grape recognition and segmentation. Due to the addition of the dual attention network in the new model, the segmentation effect of grape masks was greatly improved. And the phenomenon of recognizing overlapping grape clusters as a whole and incorrectly recognizing grape varieties was reduced. These results indicated that the new model was effective in decreasing misdetections and omissions in grape clusters. Through the above comparative analysis, the improved Mask R-CNN method was superior to the other four methods and can achieve accurate detection and segmentation of grape clusters in complex backgrounds. Based on the results shown in Figure 9 and Table 6, the Mask R-CNN and Cascade Mask R-CNN models had poor performance in segmenting grape edges, and the segmentation effect on grape cluster edges was not complete. Mask R-CNN was prone to identify overlapping grapes many times. In addition, these two models had more false detection of grape varieties. Secondly, the segmentation effect of the HTC model on grape masks was general, and there were also many false detections of grape varieties. The SOLO model performs well in grape mask segmentation, but it is also prone to false detection of grape varieties. Figure 10 shows the comparison results of these five methods in segm_mAP evaluation metrics. It shows that the improvement model proposed before 40 epochs has a lower increase in segm_mAP evaluation metrics than the other four models. As the number of epochs continues to increase, the precision metric of the improved model gradually surpasses that of the other four models and tends to stabilize around 80 epochs. The improved method proposed in this paper had greater improvements compared to the other models, not only improving the mask segmentation effect on grapes but also significantly reducing the false detection of grape types. The optimized Mask R-CNN model had the advantages of accurate detection, precise segmentation, and good segmentation quality, which was conducive to promoting the intelligent development of orchards.

Table 5. Parameter configuration of each instance segmentation model.

Method	**Backbone**	**Initial Learning Rate**	**Momentum**	**Weight Decay**	**Training Epochs**	**Batch Size**
Mask R-CNN	ResNet-50-FPN	0.02	0.9	1×10^{-4}	100	4
SOLO	ResNet-50-FPN	0.02	0.9	1×10^{-4}	100	4
Cascade Mask R-CNN	ResNet-50-FPN	0.02	0.9	1×10^{-4}	100	4
HTC	ResNet-50-FPN	0.02	0.9	1×10^{-4}	100	4
Ours	ResNet-50-FPN	0.02	0.9	1×10^{-4}	100	4

Table 6. Detection and instance segmentation results of five methods.

Method	**bbox_ map$_{0.5:0.95}$**	**bbox_ map$_{0.5}$**	**bbox_ map$_{0.75}$**	**segm_ map$_{0.5:0.95}$**	**segm_ map$_{0.5}$**	**segm_ map$_{0.75}$**	**segm_ AR**	**Parameters**
Mask R-CNN	0.891	0.976	0.955	0.806	0.974	0.938	0.852	44.01 M
SOLO	/	/	/	0.811	0.974	0.938	0.851	36.14 M
Cascade Mask R-CNN	0.894	0.970	0.953	0.791	0.972	0.927	0.832	77.05 M
HTC	0.902	0.980	0.959	0.812	0.979	0.942	0.849	77.18 M
Ours	0.905	0.982	0.963	0.821	0.982	0.959	0.861	44.4 M

The ‘/’ means that the SOLO model does not have a detection function.

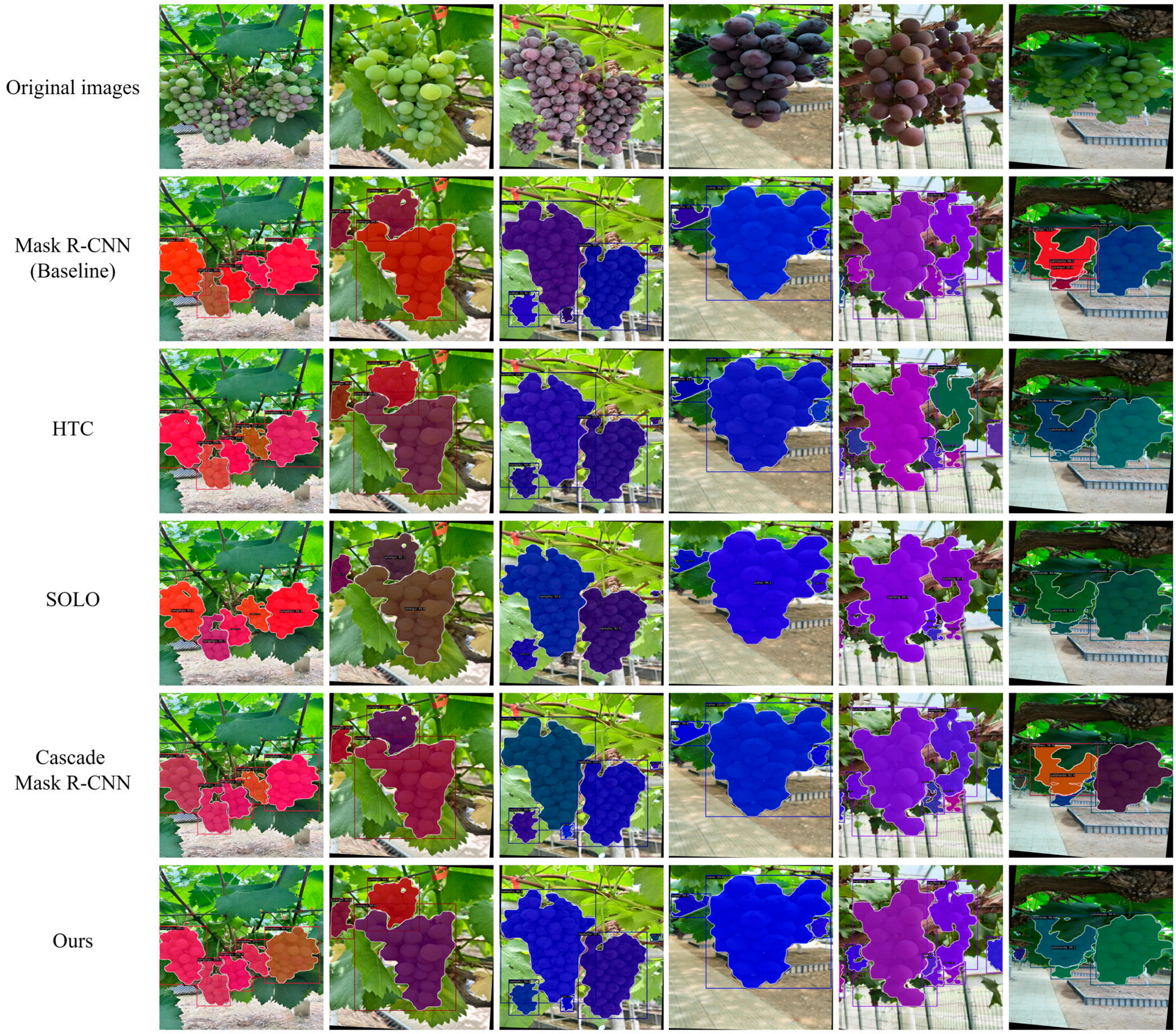

Figure 9. Grape cluster detection and segmentation results of five instance segmentation models.

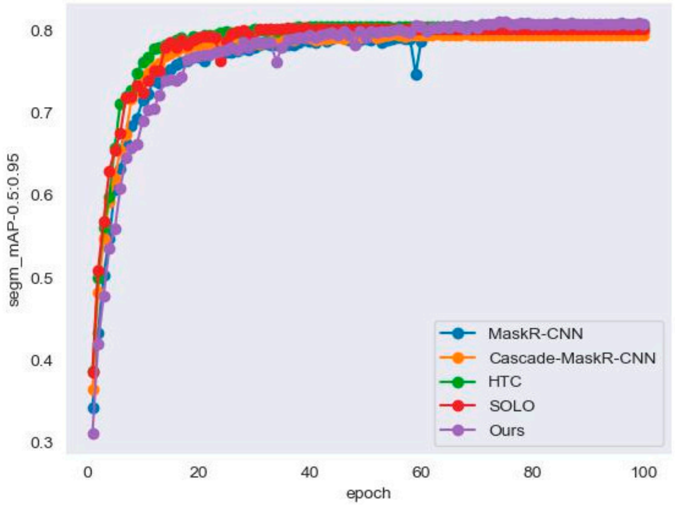

Figure 10. Comparison results of five methods in segm_mAP evaluation metric.

3.3. Statistical Analysis

As shown in Tables 3 and 4, the newly proposed optimized Mask RCNN model had greatly improved the detection, segmentation precision, and recall rate of grape clusters without significantly increasing the model parameters. In terms of the recognition precision of the eight grape varieties, some grape varieties were detected and segmented with relatively low precision, which was related to the similarity in appearance and color of the grapes. The grape color was mainly affected by grape maturity. Some grape varieties showed a variety of colors at different maturities, and some grape varieties showed a single color at different maturities. These will affect the final experimental results of the model. In the model comparison experiment and visualization results, it could be found that the proposed Mask-RCNN model was superior to other models in segmentation detection precision and identification of grape cluster varieties. This also showed that the ECA module and DANet module had good roles in promoting the task of detecting and segmenting grape clusters.

4. Discussion

4.1. Analysis of Orchard Grape Clusters Detection and Segmentation Results

The complexity of the orchard environment and the wide variety and similarity of grape varieties in the orchard brought challenges to the accurate segmentation and recognition of orchard grapes. This study proposed an improved Mask R-CNN method to solve these problems. To further improve the ability of the model to extract the characteristics of orchard grapes, the residual network combined with the ECA attention module was used. It enabled the network to better mine the subtle features of grapes, and effectively ignore the background and other irrelevant features. At the same time, a dual attention network, DANet, was added to the head of the segmentation mask to strengthen the segmentation performance of the network. It further enhanced the segmentation precision of the orchard grape cluster mask. The experimental results of the improved Mask R-CNN model show that this method can validly identify and segment grape clusters in different environments of orchards accurately.

There were also false identifications of grape species when using the improved Mask R-CNN model, as shown in Figure 11. One of the main reasons was that there were several types of grapes with high similarity. The other was that the blurred grape object in the depth of field in the picture would interfere with grape recognition. The hongdiqiu grape shown in Figure 11A was similar to the zaotian grape. Due to the factors such as the shadow of the environment itself and the similarity of the environment and fruit color, grapes belonging to the hongdiqiu variety were mistakenly detected as zaotian grapes. In the image shown in Figure 11B, the grape objects in the depth of field in the picture were relatively blurry, and there were high similarity of some grape varieties. The grapes originally belonging to the zaotian variety in the image were mistakenly detected as the jumeigui variety. Among the original grapes, the jumeigui, zaotian, and hongdiqiu grape varieties themselves had a high degree of similarity. The mean average segmentation precision of the baseline model results in Table 4 were lower than other grape varieties. In this study, although the improved Mask R-CNN method also had the phenomenon of false identification of grape varieties, comopared with the baseline model, there was a significant reduction in the occurrence of variety recognition errors.

Compared with other classic instance segmentation methods, the proposed improved Mask R-CNN model significantly improved the precision of grape cluster detection and segmentation, and its model parameter quantity was not large compared to other models. For evaluation metrics ranging from $AP_{IoU=0.5}$ to $AP_{IoU=0.75}$ and then to $AP_{IoU=0.5:0.95}$, as the model metric requirements continued to increase, the decrease in evaluation metrics of the improved Mask R-CNN model was smaller compared to other models. For the visualization results of the model test set, the improved Mask R-CNN model and some other instance segmentation can obtain a relatively complete grape cluster mask. But the improved Mask R-CNN model had better accuracy in identifying grape cluster variety than other segmentation models.

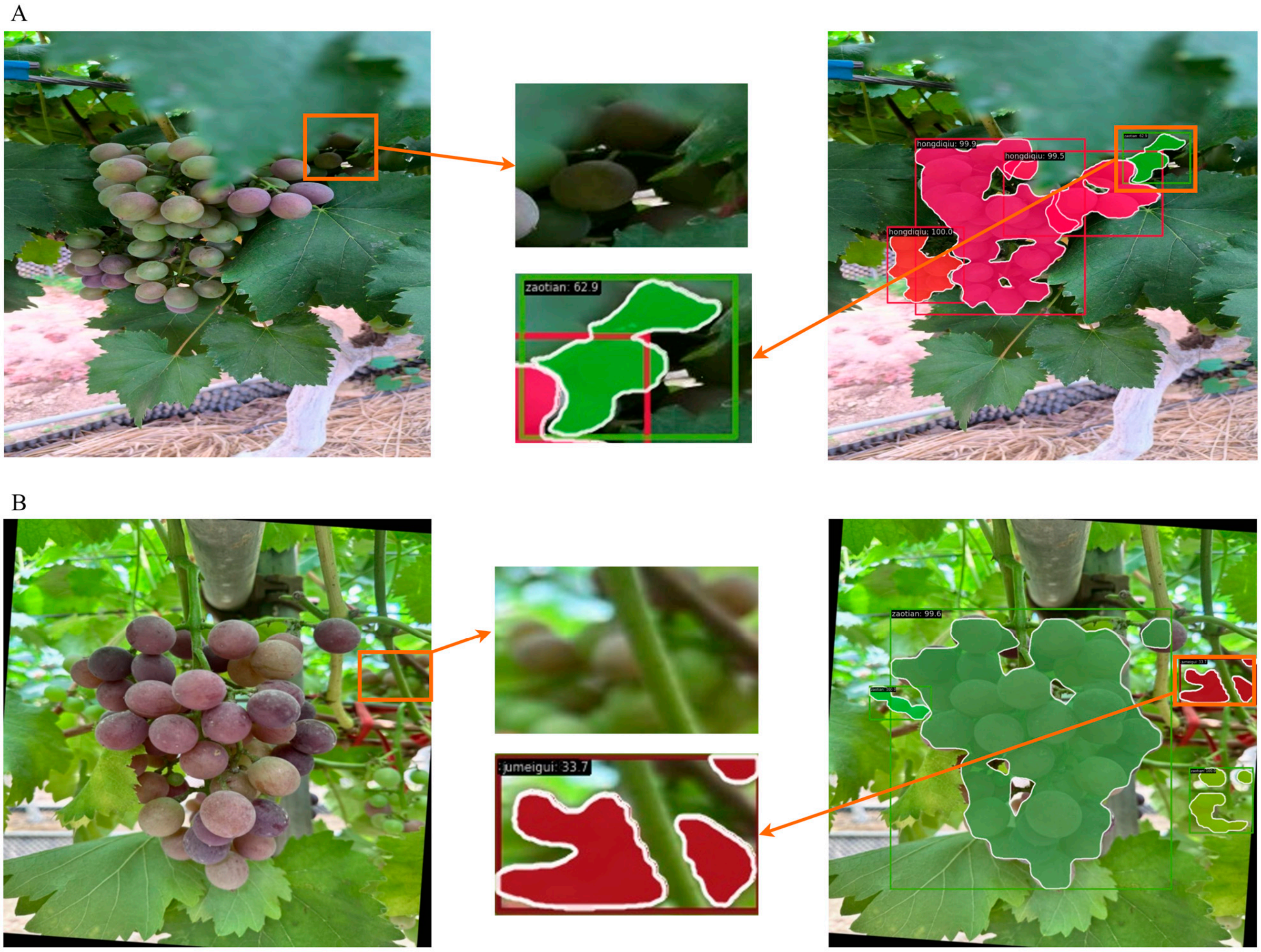

Figure 11. False detection and instance segmentation results: (**A**) Hongdiqiu grape variety error recognition image; (**B**) Zaotian grape variety error recognition image.

4.2. Effect of the Model Improvement on Grape Detection and Segmentation

An improved model was proposed by optimizing the original Mask R-CNN model. First, ResNet-50 integrated the ECA attention module combined with FPN as the backbone network. Then, the DANet module was inserted into the mask head and played a segmentation role in the original Mask R-CNN. Table 7 shows the ablation experimental results of these modules added to the original Mask R-CNN network on the test set. First, the results in Table 7 were analyzed separately from these two modules. The ECA attention module significantly improved the detection precision of bounding boxes. When the IoU was 0.75, its bounding box detection precision was improved by 1.7% compared to the baseline Mask R-CNN model. The DANet module effectively improved the segmentation effect of grape masks. When the IOU increased from 0.5 to 0.95 at 0.05 intervals, the mean average precision(mAP) for all grape varieties at different IoU thresholds was 0.816. It had increased by 1% compared with the baseline Mask R-CNN model. As can be found in Table 7, the combination of ECA and DANet was better than that of individual modules. Among the six evaluation metrics in Table 7, five metrics achieved the best results after integrating the ECA and DANet modules. The experimental results of our improved Mask R-CNN model are shown in the last row, and there is a greater improvement in the precision of bounding box detection and mask segmentation compared to the original Mask R-CNN model. For example, when the IoU was 0.75, the precision of bounding box detection and mask segmentation was improved by 0.8% and 2.1% compared to the original baseline model, respectively. This further demonstrated the effectiveness of the ECA module and DANet module in improving the Mask R-CNN model. Figure 12 shows the comparison of segm_mAP evaluation metric results in the ablation experiment of the DANet module and ECA module.

Table 7. Results of ablation experiments of Mask R-CNN model using different modules.

Method	ECA	DANet	bbox_ $\mathrm{map}_{0.5:0.95}$	bbox_ $\mathrm{map}_{0.75}$	bbox_AR	segm_ $\mathrm{map}_{0.5:0.95}$	segm_ $\mathrm{map}_{0.75}$	segm_AR
Mask R-CNN(ResNet-50)			0.891	0.955	0.925	0.806	0.938	0.852
	✓		0.898	**0.972**	0.931	0.804	0.948	0.849
		✓	0.894	0.955	0.927	0.816	0.953	0.858
	✓	✓	**0.905**	0.963	**0.933**	**0.821**	**0.959**	**0.861**

The bold numbers in Table 7 indicate the optimal results of the evaluating metrics.

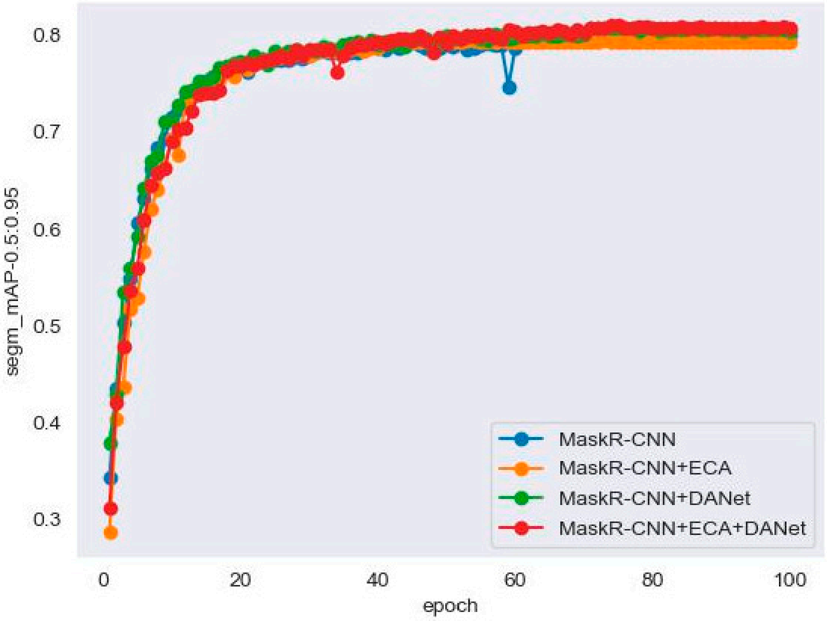

Figure 12. Comparison of segm_mAP evaluation metric results in ablation experiments of the DANet module and ECA module.

4.3. The Impact of Model Improvement Module on Other Instance Segmentation Models

Based on the existing research mentioned above, an experimental exploration of the model improvement modules on other instance segmentation models was conducted. We conducted ablation experiments on the Cascade Mask R-CNN, HTC, and SOLO network models. The above ECA attention module and DANet module were also integrated into their backbone networks and their respective mask heads. The impact of the ECA attention module and DANet module on the three instance segmentation models is shown in Table 8. To reduce computing power and adapt to the limitations of computer memory, the batch size of the three instance segmentation models was set to 1. A conclusion can be drawn from the comparison of the results in Tables 6 and 8. As the batch size was reduced from 4 to 1, the baseline detection and segmentation precision of the Cascade Mask R-CNN, HTC, and SOLO network models increased significantly. As shown in Table 8, it can be seen that after adding the ECA attention module and DANet module, the bounding box detection precision and average recall of HTC model remained unchanged compared to the original model. However, there is a minimal improvement in segmentation precision and average recall. For the Cascade Mask R-CNN model, after incorporating the ECA attention module and DANet module on the model baseline, its output evaluation metrics showed a small improvement in both bounding box detection and segmentation precision. The average recall of bounding box detection and mask segmentation also showed a small improvement. For the SOLO model, the performance was reduced after integrating the ECA attention module and the DANet module. This indicates that DANet and ECA attention modules may interfere with the original feature processing method of the SOLO algorithm and the design of the instance segmentation header. Based on the above results, it can be concluded that the ECA module and DANet module have a positive impact when integrated into the backbone network and segmentation head of the HTC model and Cascade Mask R-CNN model, respectively, while incorporating them into the SOLO model has a negative impact. This also indicated that the integration of DANet and ECA attention modules into instance segmentation models may not achieve improved results.

Table 8. Experimental results of other instance segmentation models fused ECA and DANet modules.

Method	bbox_ map$_{0.5:0.95}$	bbox_ map$_{0.75}$	bbox_AR	segm_ map$_{0.5:0.95}$	segm_ map$_{0.75}$	segm_AR
HTC	0.939	0.979	0.957	0.846	0.971	0.878
HTC + ECA + DANet	0.937	0.980	0.957	0.847	0.973	0.880
Cascade Mask R-CNN	0.930	0.972	0.948	0.837	0.963	0.869
Cascade Mask R-CNN + ECA + DANet	0.933	0.975	0.951	0.843	0.964	0.876
SOLO	/	/	/	0.826	0.940	0.960
SOLO + ECA + DANet	/	/	/	0.781	0.911	0.827

The '/' means that the SOLO model does not have a detection function.

5. Conclusions

In this study, an improved Mask R-CNN model is proposed for the accurate detection and segmentation of orchard grape clusters. Firstly, to strengthen the ability of the backbone network to extract the characteristics of orchard grapes, the ECA module was introduced into the original Mask R-CNN backbone network. The DANet module was then incorporated into the convolutional networks of the mask branches to enhance the segmentation precision of the grape cluster masks. In addition, the ECA attention module and DANet module were also used to explore their impact on other instance segmentation network models. The results showed that the idea of module fusion and replacement of network structure can effectively strengthen the performance of other instance segmentation models. Compared with the baseline Mask R-CNN model, the improved Mask R-CNN model could effectively and accurately detect and segment grape clusters of different varieties in the orchard. It also reduced the phenomenon of false detection of grape cluster varieties and showed good robustness. On the test set of the divided grape segmentation dataset, the bbox_mAP, segm_AP, bbox_AR, segm_AR, and average derivation time for each test set image were 0.905, 0.821, 0.933, 0.861, and 21.5 ms, respectively. The experimental results of different classical instance segmentation models proved that the proposed method was progressive. However, this improved Mask R-CNN model was relatively large. In addition, the precision of grape cluster mask segmentation and the correct identification of high-similarity orchard grape varieties need to be further improved. In the future, more grape varieties and grape images under different environmental conditions will be collected to further expand the dataset. Future research will be based on reducing the network model parameters, improving grape clusters detection and segmentation precision, and reducing the false detection of grape clusters with high similarity.

Author Contributions: Software, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing—Original Draft, X.H.; Investigation, Writing—Review and Editing, D.P.; Funding acquisition, Writing—Review and Editing, H.Q.; Methodology, Writing—Review and Editing, L.Z.; Resources, Conceptualization, Methodology, Visualization, Supervision, Project administration, Writing—Review and Editing, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources [grant number: 2020E10017].

Institutional Review Board Statement: Not applicable.

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

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Attri, I.; Awasthi, L.K.; Sharma, T.P.; Rathee, P. A review of deep learning techniques used in agriculture. *Ecol. Inform.* **2023**, *77*, 102217. [CrossRef]
2. Behroozi-Khazaei, N.; Maleki, M.R. A robust algorithm based on color features for grape cluster segmentation. *Comput. Electron. Agric.* **2017**, *142*, 41–49. [CrossRef]
3. Liu, S.; Whitty, M. Automatic grape bunch detection in vineyards with an SVM classifier. *J. Appl. Log.* **2015**, *13*, 643–653. [CrossRef]
4. Chauhan, A.; Singh, M. Computer vision and machine learning based grape fruit cluster detection and yield estimation robot. *J. Sci. Ind. Res.* **2022**, *81*, 866–872.
5. Wang, C.L.; Liu, S.C.; Wang, Y.W.; Xiong, J.T.; Zhang, Z.G.; Zhao, B.; Luo, L.F.; Lin, G.C.; He, P. Application of Convolutional Neural Network-Based Detection Methods in Fresh Fruit Production: A Comprehensive Review. *Front. Plant Sci.* **2022**, *13*, 868745. [CrossRef] [PubMed]
6. Naranjo-Torres, J.; Mora, M.; Hernández-García, R.; Barrientos, R.J.; Fredes, C.; Valenzuela, A. A Review of Convolutional Neural Network Applied to Fruit Image Processing. *Appl. Sci.* **2020**, *10*, 3443. [CrossRef]
7. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
8. Mohimont, L.; Roesler, M.; Rondeau, M.; Gaveau, N.; Alin, F.; Steffenel, L.A. Comparison of Machine Learning and Deep Learning Methods for Grape Cluster Segmentation. In Proceedings of the International Conference on Smart and Sustainable Agriculture, Virtual Event, 21–22 June 2021; pp. 84–102.
9. Santos, T.T.; de Souza, L.L.; dos Santos, A.A.; Avila, S. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Comput. Electron. Agric.* **2020**, *170*, 105247. [CrossRef]
10. Wang, Z.F.; Zhang, Z.H.; Lu, Y.Q.; Luo, R.; Niu, Y.; Yang, X.B.; Jing, S.X.; Ruan, C.Z.; Zheng, Y.J.; Jia, W.K. SE-COTR: A Novel Fruit Segmentation Model for Green Apples Application in Complex Orchard. *Plant Phenomics* **2022**, *2022*, 0005. [CrossRef] [PubMed]
11. Wang, C.; Yang, G.P.; Huang, Y.W.; Liu, Y.K.; Zhang, Y. A transformer-based mask R-CNN for tomato detection and segmentation. *J. Intell. Fuzzy Syst.* **2023**, *44*, 8585–8595. [CrossRef]
12. Li, Y.E.; Wang, Y.; Xu, D.Y.; Zhang, J.J.; Wen, J. An Improved Mask RCNN Model for Segmentation of 'Kyoho' (*Vitis labruscana*) Grape Bunch and Detection of Its Maturity Level. *Agriculture* **2023**, *13*, 914. [CrossRef]
13. 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; pp. 2961–2969.
14. Wang, D.D.; He, D.J. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [CrossRef]
15. Jia, W.K.; Wei, J.M.; Zhang, Q.; Pan, N.N.; Niu, Y.; Yin, X.; Ding, Y.H.; Ge, X.T. Accurate segmentation of green fruit based on optimized mask RCNN application in complex orchard. *Front. Plant Sci.* **2022**, *13*, 955256. [CrossRef] [PubMed]
16. Yu, Y.; Zhang, K.L.; Yang, L.; Zhang, D.X. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
17. Chen, S.; Song, Y.; Su, J.; Fang, Y.; Shen, L.; Mi, Z.; Su, B. Segmentation of field grape bunches via an improved pyramid scene parsing network. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 185–194. [CrossRef]
18. Shen, L.; Su, J.; Huang, R.; Quan, W.; Song, Y.; Fang, Y.; Su, B. Fusing attention mechanism with Mask R-CNN for instance segmentation of grape cluster in the field. *Front. Plant Sci.* **2022**, *13*, 934450. [CrossRef] [PubMed]
19. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
20. Jung, A. *Imgaug Documentation*; Readthedocs: Portland, OR, USA, 2019.
21. 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] [PubMed]
22. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
23. 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; pp. 770–778.

24. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
27. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
28. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 421–436.
29. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13. pp. 740–755.
30. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting objects by locations. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVIII 16. pp. 649–665.
31. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
32. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.