forests

MDPI

Article

Precious Tree Pest Identification with Improved Instance Segmentation Model in Real Complex Natural Environments

Ying Guo [1], Junjia Gao [1], Xuefeng Wang [1,*], Hongyan Jia [2], Yanan Wang [2], Yi Zeng [2], Xin Tian [1], Xiyun Mu [3], Yan Chen [1] and Xuan OuYang [1]

[1] Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China
[2] The Experimental Center of Tropical Forestry of CAF, Chinese Academy of Forestry, Pingxiang 532600, China
[3] Forestry Research Institute of Chifeng, Chifeng 024000, China
* Correspondence: xuefeng@ifrit.ac.cn

Abstract: It is crucial to accurately identify precious tree pests in a real, complex natural environment in order to monitor the growth of precious trees and provide growers with the information they need to make effective decisions. However, pest identification in real complex natural environments is confronted with several obstacles, including a lack of contrast between the pests and the background, the overlapping and occlusion of leaves, numerous variations in pest size and complexity, and a great deal of image noise. The purpose of the study was to construct a segmentation method for identifying precious tree pests in a complex natural environment. The backbone of an existing Mask region-based convolutional neural network was replaced with a Swin Transformer to improve its feature extraction capability. The experimental findings demonstrated that the suggested method successfully segmented pests in a variety of situations, including shaded, overlapped, and foliage- and branch-obscured pests. The proposed method outperformed the two competing methods, indicating that it is capable of accurately segmenting pests in a complex natural environment and provides a solution for achieving accurate segmentation of precious tree pests and long-term automatic growth monitoring.

Keywords: precious trees control; small pest segmentation; instance segmentation; Mask RCNN; Swin Transformer

Citation: Guo, Y.; Gao, J.; Wang, X.; Jia, H.; Wang, Y.; Zeng, Y.; Tian, X.; Mu, X.; Chen, Y.; OuYang, X. Precious Tree Pest Identification with Improved Instance Segmentation Model in Real Complex Natural Environments. *Forests* **2022**, *13*, 2048. https://doi.org/10.3390/f13122048

Academic Editor: Won Il Choi

Received: 17 October 2022
Accepted: 29 November 2022
Published: 1 December 2022

1. Introduction

Precious trees pest control is a global concern that is critical to ecological security and the forestry industry [1]. The accurate identification of pests in a real, complex natural environment is essential for effectively monitoring precious trees growth and providing growers with data they need to make effective decisions [2,3]. Nevertheless, manual pest identification is a time-consuming and labor-intensive process that requires specialized forestry knowledge [4,5]. Therefore, it is necessary to develop automatic and accurate precious tree pest identification methods in real field environments.

Earlier attempts at automatic pest identification centered on traditional machine learning approaches, which frequently employed standard image processing algorithms or the manual design of features and classifiers for feature extraction and object identification [6–8]. Frequently, this method employs various pest characteristics to develop an identification scheme based on images captured with the appropriate lighting and shooting angle. In a natural environment, pest identification is hampered by a number of factors, including varying illumination, overlap and occlusion of leaves, similarity in color between the pests in the larva stage and the background, and uneven color and shadows on the leaf surface. Expecting traditional algorithms to eliminate the effect of scene changes on results [9,10] is impractical. In such circumstances, conventional identification methods are often ineffective, and it is difficult to achieve superior identification performance.

Recent years have seen the successful implementation of deep learning models represented by convolutional neural networks (CNN) in a variety of computer vision-related

fields [11–15]. Due to its superior performance, it is used for a variety of pest identification tasks [16,17]. In contrast to conventional machine learning techniques, it could automatically extract high-dimensional information from training datasets, reducing the requirement for labor-intensive feature engineering.

In general, the model for pest identification based on CNN could be subdivided into classification network, detection network, and segmentation network, based on the different network structures required for different tasks [16–18]. For gathering pest information, the segmentation method, which not only finely separates the pest region but also acquires the location, category, and matching geometric properties, has proven superior to the classification and detection network approaches [19]. Pests tend to cluster and overlap in their native environment. In CNN-based object segmentation models, semantic segmentation methods accurately segment pests; however, the network generates the same mask for pests of the same class, preventing the segmentation of overlapping pests. Unlike semantic segmentation, the instance segmentation method, on the other hand, could generate a unique mask for each pest instance and differentiate between individual pests.

The mask region-based convolutional neural network (Mask RCNN) [20] is a state-of-the-art instance segmentation method that has been widely applied to a variety of segmentation and detection tasks [21,22]. These algorithms, which employ stackable learnable convolutions to capture rich information in computer vision, have proven to be highly effective. However, because to the inherent localization of convolution processes [23], CNN-based modeling of global semantic information still has limits. It is difficult to identify the small pests if only relying on the local information, as the segmentation of small pests typically relies on the comparison of local information to global background information about pests. To liberate the network from the local pattern concentration of CNNs, numerous attempts have been made to model global contextual information, with attention mechanisms being the most popular approach [24–28].

Swin Transformer uses a self-attention mechanism to capture global context information, which has demonstrated exceptional performance in multiple domains [29–35]. The self-attention mechanism assesses the output at a particular position in a sequence by concentrating on all positions and calculating the weighted average in an embedding space. In other words, the mechanism for self-attention collects contextual information from other instances. By weighting values using an attention matrix, the self-attention mechanism increases the distance or distinction between classes. Therefore, Swin Transformer automatically incorporates class relationships into its feature maps. According to a number of researchers, CNN models could be outperformed by the model combined with Swin Transformer [33]. However, previous research in this field has primarily focused on object detection [34] and semantic segmentation [35], while the challenge of small precious tree pest segmentation via instance segmentation has been addressed relatively infrequently.

Swin [36] is one of the transformer methods that utilizes hierarchical information from multi-scale feature maps and achieves superior performance across a range of vision tasks. Additionally, it generates higher-resolution feature maps than other transformer methods, which is advantageous for prediction maps containing small-scale objects. As a result, we investigated incorporating Swin Transformer into the instance segmentation framework to address the challenge of frequently small-scale precious tree pest segmentation.

This research proposed a pest instance segmentation method based on an enhanced instance segmentation framework fused with Swin Transformer, named MT, to guarantee accurate segmentation of multiple small pest individuals in complex natural environments. The particular objectives are as follows: (1) Incorporate Swin Transformer method into the backbone of the Mask RCNN to enhance the ability of the network to extract features, thereby enhancing the accuracy in precious tree pest segmentation. (2) Train and evaluate the improved Mask RCNN to achieve precise detection and segmentation of small pests in the real, complex natural environment.

The remaining sections are organized as follows: Section 2 describes and analyzes the datasets; Section 3 discusses the proposed approach and technical details; Section 4

describes and analyzes the experimental results; and Section 5 discusses conclusions and future work.

2. Materials and Methods

2.1. Image Dataset

This paper uses the larvae of *Heortia vitessoides* Moore as an example for experimental research to evaluate the effect of the proposed method on small pest identification and segmentation in real complex natural environments [37–39]. As the primary pest of the precious trees *Aquilaria sinensis* in China, the *Heortia vitessoides* Moore nibbles the leaves of *Aquilaria sinensis* in its infancy, causing the petiole to fall off [40,41], and causes severe damage to *Aquilaria sinensis* when it erupts, with a damaged plant rate of more than 90% [42,43], posing a serious threat. The larvae of *Heortia vitessoides* Moore are small and mostly group on the leaves of *Aquilaria sinensis*, and their color and texture are mostly similar to the surrounding environment, posing some identification and segmentation challenges for the model.

All images were collected in 2021 in the Tropical Forestry Experimental Center of the China Academy of Forestry Sciences, Pingxiang City, Guangxi Province. The photos were taken in a natural setting at Qingshan Forest Farm. Images were collected in natural daylight with both backlight and direct sunlight situations on sunny and cloudy days to ensure a diverse set of image samples. The images were taken with a mobile phone and saved in JPEG format with a resolution of 6240 × 4160 pixels. A total of 987 images of *Heortia vitessoides* larva were captured, including little pests, overlapping pests, pests hidden by foliage and branches, and pests with shadowing and uneven lighting on the surface, from which 198 images captured under different weather and illuminations were selected as the test set, and the remaining 798 images were used as the training set for network training. Details are presented in Table 1.

Table 1. Detailed information of training and test images.

Weather	Condition	Morphology of Pests	Number of Training Images	Number of Test Images
Sunny	Direct sunlight	little pests	55	12
		overlapping	52	11
		hidden by foliage and branches	59	13
		uneven	53	12
	Backlight	little pests	50	15
		overlapping	52	12
		hidden by foliage and branches	50	10
		uneven	55	12
Cloudy	Direct sunlight	little pests	55	12
		overlapping	52	12
		hidden by foliage and branches	55	13
		uneven	53	11
	Backlight	little pests	53	12
		overlapping	53	14
		hidden by foliage and branches	51	15
		uneven	55	12
Total			798	198

Prior to the process of pest segmentation, it was necessary to annotate a substantial quantity of images of pests, as opposed to just their categories. To accomplish this, over five forestry specialists utilized the publicly accessible annotation tool Labelme v4.9 [44] to obtain the ground truth (GT) boundaries of the visible pests in the images. GT is generated as a json file, and pixels with annotations indicate the location of pests. After labeling, the annotated images were divided 8:2 into training dataset and test dataset. Examples of the captured images and annotations are shown in Figure 1.

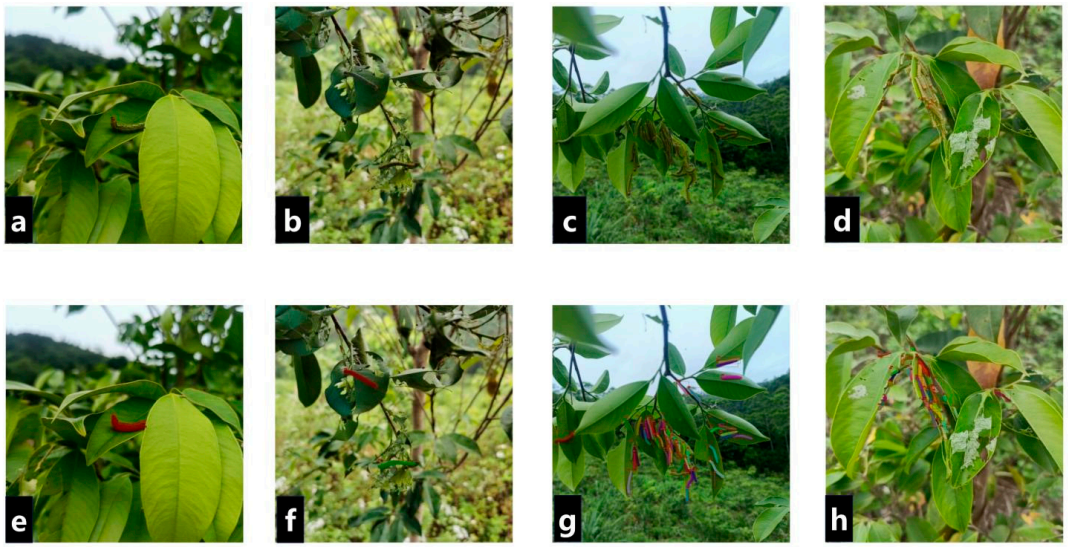

Figure 1. Examples of captured images and the corresponding annotation. (**a**) Single larva of *Heortia vitessoides* affected by shadows; (**b**) Multiple larvae of *Heortia vitessoides* affected by uneven color on the surface; (**c**) Multiple clustering larvae of *Heortia vitessoides* affected by shadows and occlusion; (**d**) Multiple clustering larvae of *Heortia vitessoides* impacted by uneven color on the surface, occlusion, and shadow; (**e**) Annotation of (**a**); (**f**) Annotation of (**b**); (**g**) Annotation of (**c**); (**h**) Annotation of (**d**).

2.2. Instance Segmentation Method of Larva Based on Improved Mask RCNN

2.2.1. Model Construction

This study suggested an enhanced Mask RCNN-based method for accurately segmenting pests in complex naturalistic environments. Modern instance segmentation methods like Mask RCNN, which extends Faster R-CNN [45] with a segmented mask generating branch, allow for accurate classification and detection. The backbone of the original Mask RCNN was suggested to be replaced with Swin Transformer model for enhancing features extraction. After receiving the outputs from the backbone, the region proposal network (RPN) generated region proposals. RoIAlign [20] gathered features from each proposal to ensure that the features were accurately aligned with the input. In the end, two operations were run simultaneously. Pest classification and regression of bounding boxes were accomplished by fully connected (FC) layers, and the fully convolutional network [46] produced highly accurate segmentation masks to identify the locations of the pests. The segmentation technique for pests based on an improved Mask RCNN network is shown in Figure 2. The following section will discuss the details.

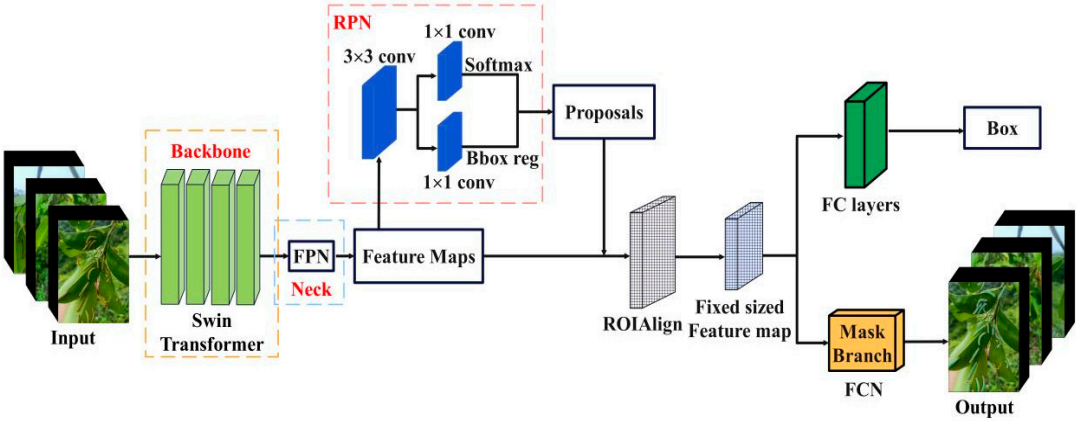

Figure 2. Framework of pest segmentation based on improved Mask RCNN.

2.2.2. Extraction of Features Based on SWIN Transformer and FPN

(1) Swin Transformer

The architecture of Swin Transformer is depicted in Figure 3. A color image is initially split into separate, non-overlapping patches by a patch splitting module [36]. The attribute of each patch is calculated by combining the raw color information of its measurement of individual pixels. Each patch is regarded as a "token." Since our implementation uses a patch size of 4 × 4, each patch has a feature dimension of 4 × 4 × 3 = 48. The raw value of this feature is projected to an undefined dimension using only a linear sequential model (denoted as C). These patch tokens have a number of Swin Transformer blocks (transformers with modified self-attention computation) affixed to them. Along with the linear embedding, Swin Transformer blocks and token count (H/4 × W/4) are referred to as "Stage 1".

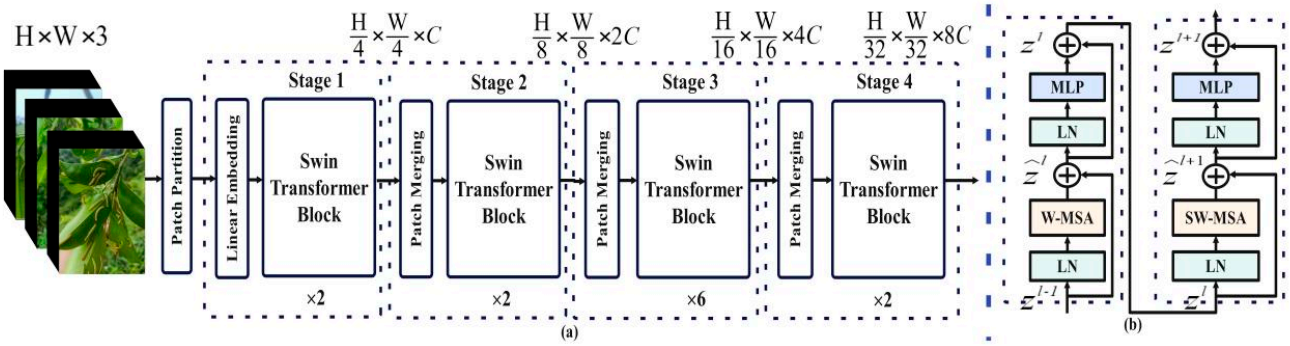

Figure 3. (**a**) The architecture of a Swin Transformer; (**b**) two successive Swin Transformer Blocks (notation presented with Equation (3)). W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively.

As the network depth increases, the number of tokens is reduced by merging patch layers to generate a hierarchical representation. The initial patch merging layer merges the features of each group of 2 × 2 adjacent patches before applying a linear layer to the fused 4C-dimensional features. This sets the output dimension to 2C and decreases the number of tokens by a factor of 2 × 2 = 4 (2 × down-sampling).

Following that, features are transformed while retaining a H/8 × W/8 resolution using Swin Transformer blocks. "Stage 2" refers to the initial stage of patch merging and feature transition. The process is then run twice, with the output resolutions being H/16 × W/16 and H/32 × W/32, respectively. As a result, the backbone networks of existing approaches for a range of visual tasks are readily replaced by the suggested architecture.

For the detail, the typical multi-head self-attention (*MSA*) module in transformer block is replaced with Swin Transformer block based on shifted windows, leaving the other layers unmodified. A shifted window-based *MSA* module, a 2-layer *MLP*, and nonlinearity are the components of Swin Transformer block. A LayerNorm (*LN*) layer and residual connection are applied before and after each *MSA* module and *MLP*, respectively. The shifting window partitioning method computes successive Swin Transformer blocks as

$$\hat{z}^l = W - MSA\left(LN(z^{l-1})\right) + z^{l-1} \quad (1)$$

$$z^l = MLP\left(LN(\hat{z}^l)\right) + \hat{z}^l \quad (2)$$

$$\hat{z}^{l+1} = SW - MSA\left(LN(z^l)\right) + z^l \quad (3)$$

$$z^{l+1} = MLP\left(LN(\hat{z}^{l+1})\right) + \hat{z}^{l+1} \quad (4)$$

where $\hat{z}^l$ and z^l are the output characteristics of the *SW-MSA* module and the *MLP* module for block l, respectively. *W-MSA* and *SW-MSA* are abbreviations for window-based

multi-head self-attention utilizing regular and shifted window partitioning configurations, respectively.

(2) Feature Pyramid Network

The pyramid hierarchy is used to build a feature pyramid with strong semantics as part of the feature pyramid network (FPN) implementation. Lin et al., (2017) indicate that FPN achieves accurate localization by utilizing high-level semantic information and dimensionality of the feature maps [47]. Figure 4 illustrates the network architecture of the FPN.

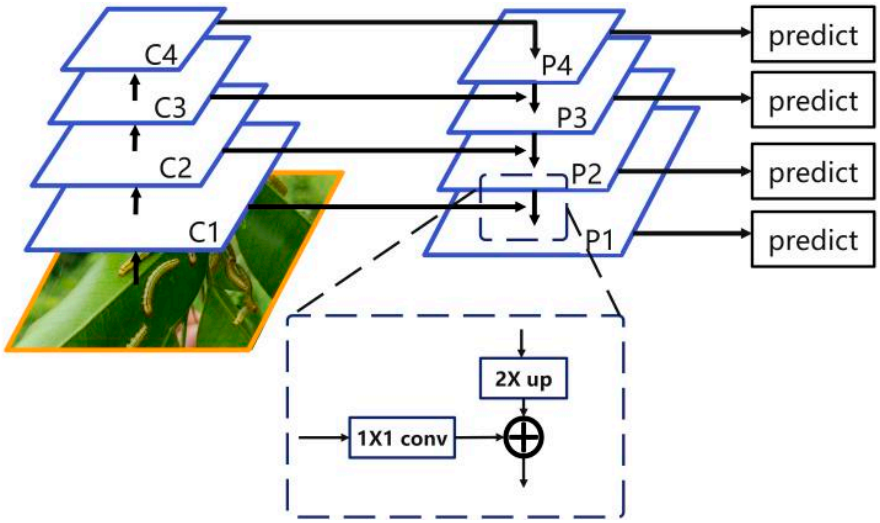

Figure 4. The network structure of FPN.

The four stages of the attended Swin Transformer correspond to the four different scales of the feature map (C1, C2, C3, and C4). The feature maps are then sent to the FPN, which generates the feature pyramid and new features [P1, P2, P3, P4]. P1 is not used in the subsequent procedures because the computation of the corresponding feature map of C1 takes a considerable amount of time. P5, which was obtained by down-sampling P4, is used in this instance.

The FPN combines the characteristics of each stage of the attended Swin Transformer, increasing network accuracy and giving the network robust semantic and spatial insight. By using the attentive Swin Transformer plus the FPN as the backbone network in this study instead of the initial backbone network, the network's capacity to extract features was improved.

2.2.3. ROI Alignment and Region of Interest (ROI) Generation

The RPN was then allowed access to the collected extracted features from the backbone network so that it could examine the ROIs for pest-infested areas. The size of the larva varies dramatically between images because of the different shooting angles. When generating the ROIs, three different area scales were created: 128 × 128, 256 × 256, and 512 × 512 based on the total number of pixels comprising pests in the captured datasets. Three aspect ratios were used: 1:2, 1:1, and 2:1. To increase the precision of the areas of interest (ROIs) that were output, nine anchors on the original image were created for each pixel on the feature map using a randomized combination of different region scales and aspect ratios. The anchors were used to predict the locations of the pests. To identify whether a target was in the front or background, the class of the ROIs was employed. Prior to generating the class and boundary values of regions of interest, the RPN performed class and boundary box regression operations. A target sequence in the foreground indicates the existence of pests within the area of interest. The bounding box was altered to exactly cover the pest-infested area using the boundary coordinates of ROIs. The regions of interest that were generated and the related feature maps were derived via RoIAlign model [20]. Using

RoIAlign, the anchor box's dimensions were adjusted to a fixed size. The retrieved features were accurately matched with the input to increase pixel-level segmentation accuracy.

2.2.4. Pest Instance Segmentation and Loss Function

RoIAlign produced feature maps, which were then fed into the convolutional and fully connected layers. The fully connected layer was used for regression analysis and classification on the bounding box, meanwhile the fully convolutional layer was employed to segment pest instances. Convolution and deconvolution were employed for instance segmentation, and the outputs of the fully connected layer were transferred to a Softmax layer for classification.

The loss function, which is essential for network training, represents the differences between the predictions and the actual data. In this work, the neural network is trained utilizing the combined loss function of the classification, mask prediction, and bounding box regression branches. The network loss is computed using Equations (5)–(8).

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{5}$$

$$L_{cls} = \sum_{i} -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \tag{6}$$

$$L_{bbox} = \frac{1}{N_{reg}} \sum_{i} p_i^* R(t_i - t_i^*) \tag{7}$$

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij}^* \log y_{ij} + (1 - y_{ij}^*) \log(1 - y_{ij}) \right] \tag{8}$$

where L_{cls} is the classification loss, L_{bbox} is the bounding box regression loss, and L_{mask} stands for the mask loss, t_i and t_i^* represent the predicted and ground truth coordinates, whereas p_i and p_i^* represent the expected probability and actual value of the anchor, respectively. The smoothing $L1$ function is $\boldsymbol{R}(\cdot)$. The mask branch output for each ROI in the enhanced Mask RCNN has a m^2 dimension, y_{ij} represents the predicted value and y_{ij}^* represents the actual value of the coordinate point (i, j) in the $m \times m$ region.

2.3. Network Training

Experiments were performed on a platform with an Intel(R) Xeon(R) CPU E5-2643 v4 processor, 96 GB of memory, and an NVIDIA Tesla K40c GPU (12 GB memory). Python 3.7 was utilized for training and testing the pest instance segmentation network on Windows 10.

The basic Mask RCNN model, which had been previously trained on the COCO dataset, was used to initialize the improved Mask RCNN in order to accelerate the training process [48]. The improved Mask RCNN network was then trained using the pest-labeling images. The following parameters were set to their respective values: 0.00001, 2.0, 0.90, 0.05, and 100 epochs for learning rate, batch size, learning momentum, weight decay, and iterations. Six hours were spent on training in its entirety.

2.4. Evaluation of the Performance of the Network Model

The performance of the suggested pest instance segmentation method was assessed using three parameters: precision, recall and $F1$ score [49]. Equations (6)–(8) can be used to calculate the parameters. The greater the values of the three parameters, the better the outcomes.

$$precision = TP/(TP + FP) \times 100\% \tag{9}$$

$$recall = TP/(TP + FN) \times 100\% \tag{10}$$

$$F1 = 2 \times precision \times recall/(precision + recall) \tag{11}$$

where TP, FP, and FN, respectively, represent true positive, false positive, and false negative.

3. Results

3.1. Instance Segmentation of Pests

To evaluate the effectiveness of the proposed method, 198 images of pests captured in a real complex natural environment were used to test the method, taken under various lighting and weather conditions. The precision, recall and F1 score were 87.2%, 90.95% and 89.0%, respectively. Examples of segmentation results and precise segmentation accuracy resulting from this approach are shown in Figure 5 and Table 2, respectively.

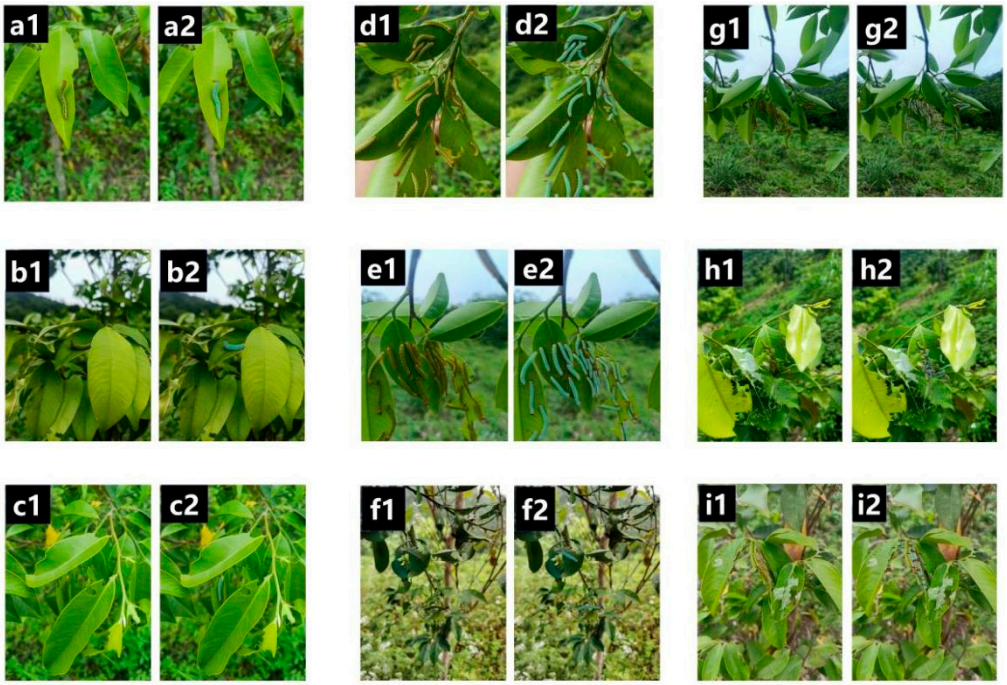

Figure 5. Examples of instance segmentation of pests: (**a1**,**a2**) Single larva of *Heortia vitessoides* and the corresponding instance segmentation result. (**b1**,**b2**) Single larva of *Heortia vitessoides* affected by shadows and the corresponding instance segmentation result. (**c1**,**c2**) Single larva of *Heortia vitessoides* sheltered by leaves and the corresponding instance segmentation result. (**d1**,**d2**) Multiple non-overlapped larvae of *Heortia vitessoides* and the corresponding instance segmentation result. (**e1**,**e2**) Multiple non-overlapped larvae of *Heortia vitessoides* affected by shadows and the corresponding instance segmentation result. (**f1**,**f2**) Multiple larvae of *Heortia vitessoides* affected by uneven color on the surface and the corresponding instance segmentation result. (**g1**,**g2**) Multiple clustering larvae of *Heortia vitessoides* affected by shadows and occlusion and the corresponding instance segmentation result. (**h1**,**h2**) Multiple clustering larvae of *Heortia vitessoides* affected by uneven color on the surface and occlusion and the corresponding instance segmentation result. (**i1**,**i2**) Multiple clustering larvae impacted by uneven color on the surface, occlusion, and shadow, as well as the related segmentation result for the instance.

Table 2. Segmentation results of pests under different conditions.

Conditions	SL	MNL	OL	SAL	MCL
precision(%)	94.85	94.4	86.6	83.9	83.3
recall(%)	96.65	97.2	89.2	88.3	87.2
F1(%)	96.0	96.0	87.9	86.0	85.2

Single larva (SL); multiple non-overlapping larvae (MNL); larvae affected by uneven color and occluded by branches and leaves (OL); larvae affected by shadow (SAL); Multiple clustering larvae (MCL).

As shown in Figure 5, the proposed method accurately segmented single larva of *Heortia vitessoides* (Figure 5a), and single larva affected by shadows and occlusion (Figure 5b,c). Figure 5d–f represents segmentation results that were satisfactory for larvae

that were non-overlapped. Our method was effective for segmenting larvae covered by branches and foliage (Figure 5b,c,h) and affected by shadows (Figure 5e,g) and uneven color on the surface(Figure 5f,i). In addition, the proposed method accurately segmented pests that were overlapping one another (Figure 5g–i).

To evaluate the segmentation outputs of larvae affected by various circumstances, the performance of the larva instance segmentation influenced by branch and foliage coverage, overlapping, uneven colors, shadowing, and poor lighting on the leaf surface was computed. (Table 2). Our method produced superior segmentation results in general. With a 96.0% F1, our method accurately segmented single larva and multiple non-overlapping larvae. F1 values for larvae affected by uneven color and occluded by branches and leaves were 87.9%. This suggests that our method could successfully mitigate the effect of color on segmentation outcomes. Our method could effectively and accurately segment larvae with shadows, as evidenced by *F*1 values of 86.0%. Our method was robust for segmenting overlapped and occluded larvae, as evidenced by *F*1 values of 85.2% for larvae occluded by branches and leaves, respectively.

3.2. Comparison with Other Instance Segmentation Methods

Precision, recall, and *F*1 scores were used to assess the effectiveness of the improved Mask RCNN-based pest instance segmentation approach, and the performance of the approach was contrasted to that of Mask RCNN with ResNet50 (MR50), and ResNet101 (MR101). All three networks were trained and assessed using the same test, validation, and training sets. Mask RCNN was trained with the following parameters: 0.02, 2.0, 0.90, 0.0001, 100 epoch, learning rate, batch size, learning momentum, and weight decay. The segmentation outcomes for the four techniques used on the test set are displayed in Table 2.

In terms of precision, recall, and *F*1, as shown in Table 3, our method performed better than the other methods. Our approach had an accuracy of 87.23%, which was 10.78% and 10.03% greater than MR50 and MR101, respectively. The recall of our approach was 90.95%, which was greater than both MR50 and MR101. The proposed method was more accurate in terms of the *F*1 score than the alternative methods (89.03%). It outperformed the MR50 by 14.33% and the MR101 by 14.38%. Based on the comparison results, it is feasible to conclude that the approach for pest instance segmentation proposed in this work, based on the enhanced Mask RCNN, could segment larvae in a real complex natural environment efficiently and precisely.

Table 3. Comparison with Mask RCNN methods.

Method	Precision (%)	Recall (%)	F1 (%)
MT	87.23	90.95	89.03
MR50	76.45	79.90	74.70
MR101	77.20	81.30	74.65

4. Discussion

4.1. Analysis of the Results of Pest Segmentation

The initial backbone of the Mask RCNN model was replaced with a Swin Transformer model to better boost segmentation performance. This allowed the approach to concentrate on pertinent pest characteristics in imagery while ignoring background features. This increased both the segmentation accuracy and the network's capacity to retrieve characteristics from larvae. It is possible to draw the conclusion that the suggested method may successfully separate pests in datasets based on an analysis of the segmented images of the enhanced Mask RCNN-based method.

The results of the experiment revealed that the clustering larvae obscured by branches and leaves, as well as affected by shadows, had lower performance than the majority of the other evaluated conditions (Table 2). This may be due to differences in the factors that contribute to the situation. Variable lighting and shadows on the surface of overlapping

larvae are considered to have further contributed to the phenomenon. However, the training set was unable to cover every possible situation, resulting in a slightly lower recall rate for these larvae. The extensive similarity between the ground color and the background color may affect the accuracy of segmentation for green larvae. The training dataset will be expanded in the future, and the performance of the model to extract features will be improved. The majority of pests impacted by these inclusion conditions were segregated accurately, despite recall values for clustering larvae hidden by branches and leaves being slightly lower.

As illustrated in Figure 6, mis-segmentation was discovered when our approach was employed to segment pests. In most cases, mis-segmentation was caused by diminutiveness of the pests depicted in the images. The results of Figure 6a indicated that the larvae of which only the heads were recorded in the photos were always missed. Another situation that could easily lead to mis-segmentation is when the background color is significantly similar to that of multiple pests, as shown in Figure 6b. Despite the existence of mis-segmentation, our method was capable of achieving optimal segmentation accuracy for all images.

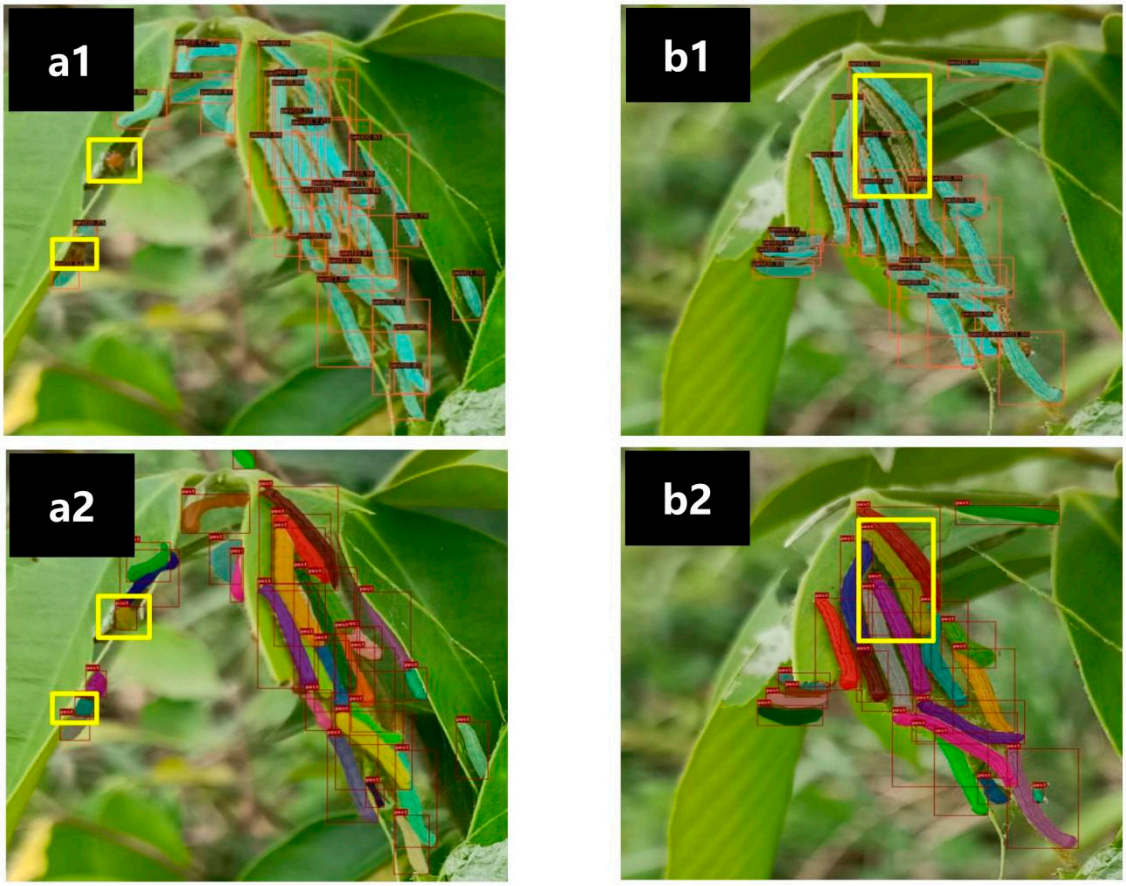

Figure 6. Examples of mis-segmentation: (**a1**) Segmentation result of only the heads of *Heortia vitetiae* larvae caught in images; (**a2**) The ground truth of (**a1**); (**b1**) The segmentation result for cases in which the backdrop color is comparable to that of multiple pests; (**b2**) The ground truth of (**b1**).

4.2. Results of Adding SWIN Transformer to Pest Segmentation

The method that was suggested was an enhanced modification of Mask RCNN. To enhance the feature extraction performance of Mask RCNN on pest segmentation, Swin Transformer model replaced the basic network backbone. The segmentation results with various structures and the accompanying model parameters were evaluated on the test set in order to assess the impact of Swin Transformer component on the functionality of the network model. Tables 3 and 4 illustrate the corresponding outcomes.

Table 4. Comparison of MR50, MR101, and MT parameters.

Method	Model Size (MB)	GFLOPs	Parameters
MT	180	135.38	47.37
MR50	168	329.33	43.75
MR101	240	481.48	62.74

The network model parameters (Table 4) showed that the magnitude and number of the model's parameters increased compared to MR50, while the network computation work was appropriately reduced due to the replacement by Swin Transformer model. Based on the *F*1 score, precision, and recall of the Mask RCNN replaced by Swin Transformer, it is clear that the segmentation accuracy improved significantly after the addition of Swin Transformer model, indicating that the Mask RCNN with Swin Transformer as its backbone improved pest segmentation accuracy. The segmentation accuracy improved as the structure and layers of Swin Transformer became more complex. Despite increasing the model size of the more complex structure and layers, pest segmentation accuracy was optimized.

5. Conclusions

In this study, an improved Mask RCNN was developed for the accurate instance segmentation of pests in a realistic, complicated natural environment. On the basis of the Mask RCNN network, a network model incorporating an attention mechanism was developed in order to enhance the feature extraction capability of the backbone network. The network was constructed by incorporating deformable convolution and Swin Transformer attention module into the backbone network of the original Mask RCNN. In comparison to the original Mask RCNN, the improved network model showed stronger segmentation capabilities. The pest instance segmentation approach based on the improved Mask RCNN segmented small pests, overlapping pests, pests obscured by foliage and branches, and pests with shadows and uneven lighting on the surface effectively and precisely. The precision, recall, and F1 score for the approach were 87.23%, 90.95%, and 89.01%, respectively. This approach accurately segmented pests under various weather and shooting situations in a complex natural environment. In general, the proposed method could significantly improve the segmentation performance of Mask RCNN, which may be useful for the effective treatment of pests that affect precious trees. In the future, we will collect images of other pests, expand our database of pests in a variety of situations, and explore methods to further optimize the networking structure and enhance segmentation accuracy.

Author Contributions: Conceptualization, Y.G. and X.W.; methodology, Y.G.; software, J.G.; validation, Y.C., X.O. and X.M.; investigation, H.J., Y.W. and Y.Z.; resources, X.T.; data curation, Y.G. and J.G.; writing—original draft preparation, Y.G. and J.G.; writing—review and editing, X.W.; funding acquisition, Y.G. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Special Funds for Fundamental Research Business Expenses of the Central Public Welfare Research Institution's "Study on Image Diagnosis Technology of Main Diseases and Insect Pests of Rare Tree Species", grant number CAFYBB2021ZB002; "Study on classification of tree species based on cooperative multi optical sensors", grant number CAFYBB2022SY030; "Study on key technologies of forest resources output", grant number CAFYBB2021SY006 and the Zhejiang Provincial Academy Cooperative Forestry Science and Technology Project "Research and application of forest resources monitoring technology in zhejiang province based on sky earth integration", grant number 2020SY02.

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

Acknowledgments: We would like to thank the Tropical Forestry Experimental Center of the China Academy of Forestry Sciences for capturing the valuable datasets.

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

References

1. Waters, W.E.; Stark, R.W. Forest pest management: Concept and reality. *Annu. Rev. Entomol.* **1980**, *25*, 479–509. [CrossRef]
2. Ding, W.; Taylor, G. Automatic moth detection from trap images for pest management. *Comput. Electron. Agric.* **2016**, *123*, 17–28. [CrossRef]
3. Sun, Y.; Liu, X.; Yuan, M.; Ren, L.; Wang, J.; Chen, Z. Automatic in-trap pest detection using deep learning for pheromone-based Dendroctonus valens monitoring. *Biosyst. Eng.* **2018**, *176*, 140–150. [CrossRef]
4. Zhao, Y.; Gong, L.; Huang, Y.; Liu, C. A review of key techniques of vision-based control for harvesting robot. *Comput. Electron. Agric.* **2016**, *127*, 311–323. [CrossRef]
5. Marini, L.; Ayres, M.P.; Jactel, H. Impact of Stand and Landscape Management on Forest Pest Damage. *Annu. Rev. Entomol.* **2022**, *67*, 181–199. [CrossRef]
6. Solis-Sánchez, L.O.; Castañeda-Miranda, R.; García-Escalante, J.J.; Torres-Pacheco, I.; Guevara-González, R.G.; Castañeda-Miranda, C.L.; Alaniz-Lumbreras, P.D. Scale invariant feature approach for insect monitoring. *Comput. Electron. Agric.* **2011**, *75*, 92–99. [CrossRef]
7. Xia, C.; Lee, J.-M.; Li, Y.; Chung, B.-K.; Chon, T.-S. In situ detection of small-size insect pests sampled on traps using multifractal analysis. *Opt. Eng.* **2012**, *51*, 1–13. [CrossRef]
8. Ebrahimi, M.; Khoshtaghaza, M.H.; Minaei, S.; Jamshidi, B. Vision-based pest detection based on SVM classification method. *Comput. Electron. Agric.* **2017**, *137*, 52–58. [CrossRef]
9. Tsaftaris, S.A.; Minervini, M.; Scharr, H. Machine learning for plant phenotyping needs image processing. *Trends Plant Sci.* **2016**, *21*, 989–991. [CrossRef]
10. Fuentes, A.; Yoon, S.; Park, D.S. Deep learning-based techniques for plant diseases recognition in real-field scenarios. In Proceedings of the International Conference on Concepts for Intelligent Vision Systems, Auckland, New Zealand, 10–14 February 2020; Volume 12002, pp. 3–14.
11. Yang, D.; Li, S.; Peng, Z.; Wang, P.; Wang, J.; Yang, H. MF-CNN: Traffic flow prediction using convolutional neural network and multi-features fusion. *IEICE Trans. Inf. Syst.* **2019**, *102*, 1526–1536. [CrossRef]
12. Sundararajan, S.K.; Sankaragomathi, B.; Priya, D.S. Deep belief CNN feature representation based content based image retrieval for medical images. *J. Med. Syst.* **2019**, *43*, 1–9. [CrossRef]
13. Melnyk, P.; You, Z.; Li, K. A high-performance CNN method for offline handwritten Chinese character recognition and visualization. *Soft Comput.* **2020**, *24*, 7977–7987. [CrossRef]
14. Li, J.; Mi, Y.; Li, G.; Ju, Z. CNN-based facial expression recognition from annotated rgb-d images for human–robot interaction. *Int. J. Hum. Robot.* **2019**, *16*, 1941002. [CrossRef]
15. Kumar, S.; Singh, S.K. Occluded thermal face recognition using bag of CNN ($ Bo $ CNN). *IEEE Signal Processing Lett.* **2020**, *27*, 975–979. [CrossRef]
16. Li, R.; Wang, R.; Xie, C.; Liu, L.; Zhang, J.; Wang, F.; Liu, W. A coarse-to-fine network for aphid recognition and detection in the field. *Biosyst. Eng.* **2019**, *187*, 39–52. [CrossRef]
17. Thenmozhi, K.; Reddy, U.S. Crop pest classification based on deep convolutional neural network and transfer learning. *Comput. Electron. Agric.* **2019**, *164*, 104906. [CrossRef]
18. Qiao, Y.; Truman, M.; Sukkarieh, S. Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming. *Comput. Electron. Agric.* **2019**, *165*, 104958. [CrossRef]
19. Li, W.; Zheng, T.; Yang, Z.; Li, M.; Sun, C.; Yang, X. Classification and detection of insects from field images using deep learning for smart pest management: A systematic review. *Ecol. Inform.* **2021**, *66*, 101460. [CrossRef]
20. 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.
21. Khan, M.A.; Akram, T.; Zhang, Y.-D.; Sharif, M. Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. *Pattern Recognit. Lett.* **2021**, *143*, 58–66. [CrossRef]
22. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
23. Cordonnier, J.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. In Proceedings of the The International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
24. 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, 15–20 June 2019; Volume 3146, pp. 3146–3154.
25. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
26. Fu, J.; Liu, J.; Jiang, J.; Li, Y.; Bao, Y.; Lu, H. Scene segmentation with dual relation-aware attention network. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2547–2560. [CrossRef] [PubMed]
27. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
28. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.

29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; p. 30.
30. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
31. Li, H.; Yang, F.; Zhao, Y.; Xing, X.; Zhang, J.; Gao, M.; Huang, J.; Wang, L.; Yao, J. DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Volume 12908, pp. 206–216.
32. Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; Zheng, Y. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Volume 12908, pp. 45–54.
33. Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Processing Syst.* **2021**, *34*, 2136–2147.
34. Xu, Y.; Zhu, J.-Y.; Chang, E.; Tu, Z. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 964–971.
35. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.-C. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Volume 12349, pp. 108–126.
36. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
37. Jiang, K.B.; Zhu, W.J.; Pan, W.; He, Z.D.; Zhu, B.Z. Genetic Diversity Analysis of Agarwood Agarwood Based on SRAP Markers. *J. Cent. South Univ. For. Technol.* **2020**, *40*, 131–136.
38. Pang, S.J.; Zhang, P.; Yang, B.G.; Liu, S.L.; Deng, S.K.; Feng, C.L. Effects of gap size on the growth and development of artificially regenerated saplings of Agarwood agarwood. *J. Northwest AF Univ.* **2020**, *48*, 83–88.
39. Zhang, X.X. Research progress on the development and utilization of Agarwood. *Shelter. For. Sci. Technol.* **2020**, *4*, 63–66.
40. Song, X.C.; Wang, X.Y.; Yang, G.; Huang, G.H.; Zhou, Z.Z.; Liang, K.N.; Zhang, Q.Q. Induction of Agarwood incense formation by mixing inorganic salts and hormones. *For. Sci.* **2020**, *56*, 121–130.
41. Hong, R.H.; Yin, J.F.; Chen, Y.; Xu, J.H.; Huang, X.Q. Research progress on the important pest of Pseudomonas japonica. *Trop. For.* **2019**, *47*, 66–68.
42. Wang, Z.; Xie, W.Z.; Zhu, C.Q.; Lu, X.L.; Cao, C.L.; Wen, X.J. Emergence and reproductive behavior rhythm of the yellow leaf borer. *China For. Dis. Insects* **2018**, *37*, 24–27.
43. Mao, Y.T.; Zhang, M.; Jin, X.F.; Ma, T.; Wang, C.; Sun, Z.H.; Chen, X.Y.; Li, Y.Z.; Wen, X.J. Study on the resistance of Agarwood vulgaris to Yellow leaf borer. *J. South China Agric. Univ.* **2017**, *38*, 89–96.
44. Torralba, A.; Russell, B.C.; Yuen, J. Labelme: Online image annotation and applications. *Proc. IEEE* **2010**, *98*, 1467–1484. [CrossRef]
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; p. 28.
46. 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; pp. 3431–3440.
47. 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.
48. 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 European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
49. Tian, Y.; Yang, G.; Wang, Z.; Li, E.; Liang, Z. Instance segmentation of apple flowers using the improved mask R–CNN model. *Biosyst. Eng.* **2020**, *193*, 264–278. [CrossRef]