Spruce Counting Based on Lightweight Mask R-CNN with UAV Images

Wenjing Zhou 1,2, Xueyan Zhu 1,2, Mengmeng Gu 3, Fengjun Chen 1,2 *
[1] College of engineering, Beijing Forestry University, Beijing, 100083, China.
[2] Beijing Laboratory of urban and rural ecological environment, Beijing, 100083, China.
[3] Department of Horticultural Science, Texas A&M University, College Station TX, 77843, USA.

Received: January 25, 2021. Revised: July 2, 2021. Accepted: July 15, 2021. Published: July 20, 2021.

***Abstract*—To achieve rapid and accurate counting of seedlings on mobile terminals such as Unmanned Aerial Vehicle (UAV), we propose a lightweight spruce counting model. Given the difficulties of spruce adhesion and complex environment interference, we adopt the Mask R-CNN as the basic model, which performs instance-level segmentation of the target. To successfully apply the basic model to the mobile terminal applications, we modify the Mask R-CNN model in terms of the light-weighted as follows: the feature extraction network is changed to MobileNetV1 network; NMS is changed to Fast NMS. At the implementation level, we expand the 403 spruce images taken by UAV to the 1612 images, where 1440 images are selected as the training set and 172 images are selected as the test set. We evaluate the lightweight Mask R-CNN model. Experimental results indicate that the Mean Counting Accuracy (MCA) is 95%, the Mean Absolute Error (MAE) is 8.02, the Mean Square Error (MSE) is 181.55, the Average Counting Time (ACT) is 1.514 s, and the Model Size (MS) is 90Mb. We compare the lightweight Mask R-CNN model with the counting effects of the Mask R-CNN model, the SSD+MobileNetV1 counting model, the FCN+Hough circle counting model, and the FCN+Slice counting model. ACT of the lightweight Mask R-CNN model is 0.876 s, 0.359 s, 1.691 s, and 2.443 s faster than the other four models, respectively. In terms of MCA, the lightweight Mask R-CNN model is similar to the Mask R-CNN model. It is 4.2%, 5.2%, and 9.3% higher than the SSD+MobileNetV1 counting model, the FCN+Slice counting model, and the FCN+Hough circle counting model, respectively. Experimental results demonstrate that the lightweight Mask R-CNN model achieves high accuracy and real-time performance, and makes a valuable exploration for the deployment of automatic seedling counting on the mobile terminal.**

***Keywords*—Spruce counting, Unmanned Aerial Vehicle, Lightweight, Mask R-CNN**

I. Introduction

Seedling inventory is an important standard for nursery stock companies to carry out cost accounting, seedling pricing, and profit forecasting. Accurate and rapid seedling stock counting is an urgent production problem that nursery stock companies need to solve [1]. At present, the counting of seedlings still mainly depends on labor, which has problems such as high cost, slow data update, and unstable accuracy, and the counting task of seedlings has the characteristics of mootonous, high frequency, and high precision requirements. Therefore, it is an ideal solution to quickly and accurately count the number of seedlings by using Unmanned Aerial Vehicle (UAV) and other mobile devices combined with computer vision technology [2-4].

Seedling counting is a task of target recognition in essence. This paper mainly counts spruce images taken by UAVs to realize counting automation. At present, the methods of target recognition are mainly divided into two categories: traditional machine vision method and deep learning method. Recognition methods based on traditional machine vision have been applied to count mangoes [5], wheat [6], etc. These methods generally train BP neural networks, support vector machines, etc. after extracting shallow features such as color [7-9], texture [10-11], and shape [8]. R. Linker et al. used color and smoothness to detect high-probability pixels belonging to apples [12], connected these pixel sets, and the contours of these seed regions are divided into arcs and amorphous segments; then the combination of these arcs and the resulting circle is compared with a simple apple model. The correct rate of counting apples under light conditions exceeds 85%, but it leads to a large number of false-positive detections; the correct rate of images under under-exposure conditions is close to 95%, and the false detection rate is less than 5%. Because traditional machine vision methods are difficult to overcome the interference of complex environments, in natural environments, the effect of using machine vision methods for recognition and counting tasks is not very good. Due to the limited representation capabilities of shallow features such as color, texture, and shape, the recognition methods based on deep learning can automatically extract shallow features and deep features, then recognize and counts targets. Chen et al. tried to use FCN to segment apples and oranges, then used CNN to count the number of apples and oranges in the connected area [13-14]. The accuracy rates under natural light conditions were 91.3% and 96.8%, respectively. Comparing the two counting methods for apple, it can be seen that the deep learning method is more robust to the interference of complex environments. It improves accuracy and simplifies the complicated counting process. Therefore, the application methods based on deep learning have become the research hotspot of scholars at home and abroad, the counting methods based on deep learning have been applied to oil palm [15], spruce [16], and tomato [17].

Current research has shown that the recognition method based on deep learning has high accuracy and strong robustness to interferences such as target adhesion, weed background, and light shadows. For individual spruce identification, the main difficulty is spruce adhesion and complex background interference, so the deep learning method is suitable for the research of seedling counting. After analysis and comparison, the recognition method based on deep learning has high accuracy and strong robustness to interferences such as target adhesion, weed background, and light shadows. This kind of method is suitable for the study of seedling counting. One of the advantages of Mask R-CNN is that it combines the characteristics of target detection and segmentation [18-19]. Another advantage of using Mask R-CNN is that it avoids the problem of complex background and plant adhesion. The instance segmentation model Mask R-CNN has been applied to the segmentation of flowers [20], and trees [21]. Xu et al. used Mask R-CNN to classify and count pasture livestock images collected by the quadcopter [22], and finally obtained a livestock classification accuracy rate of up to 96% and a livestock count accuracy rate of 92%. Jiang et al. used the Mask R-CNN model for weed detection [23], and the accuracy of weed recognition in field trials reached 91%. Based on the characteristics that can segment the target at the instance level, Mask R-CNN effectively solves the adhesion between spruce plants and has good robustness to complex backgrounds. Therefore, we adopt the Mask R-CNN model as the basic framework for seedling counting.

Given the time-consuming and labor-intensive status of seedling counting, this paper takes spruce as the research object, according to the characteristics of Mask R-CNN that can accurately detect and segment objects, it is adopted as the basic model of spruce counting. To apply to mobile devices such as UAVs, the basic model is innovatively improved [24]. The first step in the improvement process is to replace the feature extraction network with MobileNetV1 to reduce the parameter quantum [25]; the second step used to increase the calculation speed is to replace NMS with Fast NMS [26-27]. We solve the key algorithm design problem for counting seedlings with mobile devices such as UAVs.

II. Materials

A. Image acquisition and expansion

In September 2019, at the Inner Mongolia Nursery Base (111°49'47"E, 40°31'47"N, altitude 1 134m), spruce images were collected by using the DJI Phantom 4 UAV. We make a shooting plan according to the actual situation [28], the shooting angle is -90°, the image size is 4000×3000 pixels, and the shooting time is 5:30-7:30, 9:00-10:00 and 17:00-18:00, to adapt to the influence of different lighting conditions on image acquisition during the day. Before shooting, it is necessary to confirm whether the power and memory capacity of the UAV meet the shooting requirements. After the UAV taking off, we need to aim its lens at the ground and control it to traverse the nursery in an S-shaped trajectory to collect the spruce orthographic projection image, the flying height is set to 8m, 10m, 12m, 15m, and so on. We measure environmental parameters such as light intensity, wind speed, humidity, and temperature during the flight. Using UAVs to collect images is usually affected by light, shadow, wind, etc. [29], cause phenomena such as overexposure, underexposure, image shift, blur. The effects of shooting angle and flying height can also cause changes in the position and size of plants in the collected images. Therefore, we use the data enhancement methods shown in Figure 1 to expand the data set and improve the versatility of the model. In the end, 1612 images were obtained, of which 1440 images in the training set and 172 images in the test set.

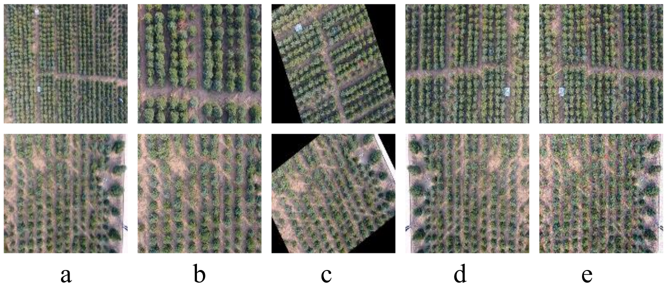

Figure. 1 Spruce pictures after data enhancement. (a)Original image. (b) Randomly cropped image. (c) Randomly rotated image. (d) Randomly flipped image. (e) Image with random noise.

Since the spruce canopy in the image taken by UAV is approximately elliptical, we change the rectangular label box of LabelImg to a circular label box to reduce visual errors, define each spruce in the image with center coordinates (x, y) and radius r, and save it as an XML file after marking. To simplify the labeling process, the model automatically converts the XML file into a mask file in the training process, which will improve the training efficiency of the spruce counting model. The mask image automatically generated from the XML file is shown in Figure 2.

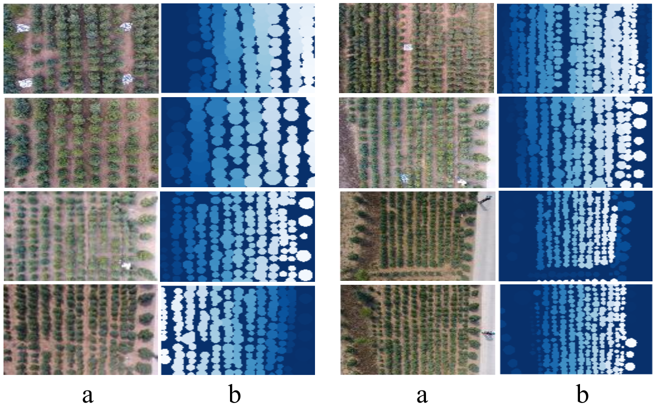

Figure. 2 Automatically generated mask image. (a)Original image. (b)Mask image.

B. Feature analysis

The spruce image obtained by UAV is shown in Figure 3. The adhesion of the spruce canopy leads to blurred edges between adjacent spruces, making counting more difficult. Therefore, solving the problem of adhesion between plants is the primary task of spruce counting. Previous studies of traditional image processing technology have not dealt with the disturbance from the complex background. To count the spruce in the image taken in the outdoor natural environment, we must first pay attention to the influence of adhesion.

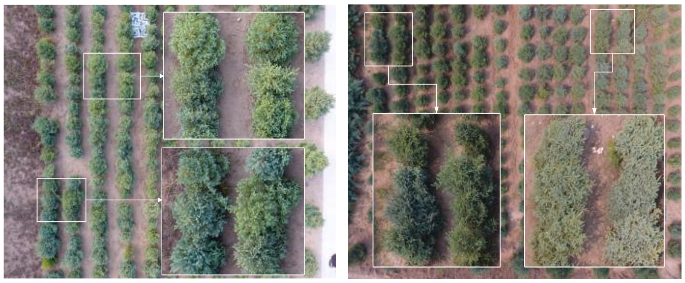

Figure. 3 Images of spruce taken by UAV.

III. METHODS

A. *Mask R-CNN model*

The Mask R-CNN model adds a mask branch for predicting segmentation to the region of interest, and the mask branch is parallel to the original classification and bounding box regression of the model. So it can perform instance segmentation for different individuals of the same category. We try to use the Mask R-CNN model as the basic model to solve the adhesion problem and then count the spruce. The model is mainly composed of feature extraction network, Region Proposal Network (RPN), and Fully Convolutional Network (FCN). Firstly, the image is input into the Mask R-CNN model, and the feature map is obtained by extracting the target features through the feature extraction network (e.g., ResNet50 [30], ResNet101, and ResNeXt101 [31], etc.). Then we will set k Regions of Interest (ROIs) for each point on the feature map to obtain multiple candidates ROIs, which are sent to the RPN network. RPN network will classify the foreground and background and return the bounding box to complete the filtering of some candidate ROI. Finally, when ROIAlign on the filtered ROI is completed, perform classification, bounding box regression and mask generation on the final ROI.

B. *Lightweight Mask R-CNN model*

One major drawback of Mask R-CNN is that it has a complex structure, a large amount of calculation, and slow calculation speed, so it is not competent for the counting work on the mobile terminal. The study began with two targeted improvements that achieve the purpose of simplifying the model, reducing the amount of calculation, and increasing the speed. The lightweight model MobileNetV1 is adopted as the feature extraction network, and Fast NMS is applied instead of NMS to construct a lightweight Mask R-CNN spruce counting model.

1) *Feature extraction network MobileNetV1*

The MobileNetV1 network mainly adopts deep separable convolution instead of the standard convolution of the VGG network [32]. To reduce the parameter quantum, the MobileNetV1 uses the width scaling factor. These characteristics determine that it has a small size and few parameters and is suitable for mobile terminals. The input image size of the network is 224×224 pixels, and its main structure is shown in Table 1 [33].

Table. 1 MobileNetV1 body architecture

Number of network layers	**Types**	**Stride**	**Filter shape**	**Input size**
1	Conv	2	3×3×3×32	224×224×3
2	Conv dw	1	3×3×32	112×112×32
3	Conv	1	1×1×32×64	112×112×32
4	Conv dw	2	3×3×64	112×112×64
5	Conv	1	1×1×64×128	56×56×64
6	Conv dw	1	3×3×128	56×56×128
7	Conv	1	1×1×128×128	56×56×128
8	Conv dw	2	3×3×128	56×56×128
9	Conv	1	1×1×128×256	28×28×128
10	Conv dw	1	3×3×256	28×28×256
11	Conv	1	1×1×256×256	28×28×256
12	Conv dw	2	3×3×256	28×28×256
13	Conv	1	1×1×256×512	14×14×256
14~18	5×Conv dw	1	3×3×512	14×14×512
19~23	5×Conv	1	1×1×512×512	14×14×512
24	Conv dw	2	3×3×512	14×14×512
25	Conv	1	1×1×512×1024	7×7×512
26	Conv dw	2	3×3×1024	7×7×1024
27	Conv	1	1×1×1024×1024	7×7×1024
28	Avg Pool	1	Pool 7×7	7×7×1024
29	FC	1	1024×1000	1×1×1024
30	Softmax	1	Classifier	1×1×1000

Among them, Conv dw is depthwise convolution, which is used for feature extraction and filtering; Conv is pointwise convolution, which is used to connect different channels to adjust the number of output channels. The first to 27 layers of the MobileNetV1 main architecture are used to construct the lightweight Mask R-CNN backbone feature extraction network. When the input image size is 224×224 pixels, the image size is compressed on the 1, 4, 8, 12, and 24 layers of the network, the output feature maps are taken out and then input to the Region Proposal Network (RPN). Depthwise separable convolution achieves the decoupling of channel correlation and spatial correlation by decomposing conventional convolution into two parts: depthwise convolution and pointwise convolution. Among them, depthwise convolution filters the channels of the input image, and pointwise convolution connects different channels. The basic structure of the depthwise separable convolution is shown in Figure 4[34].

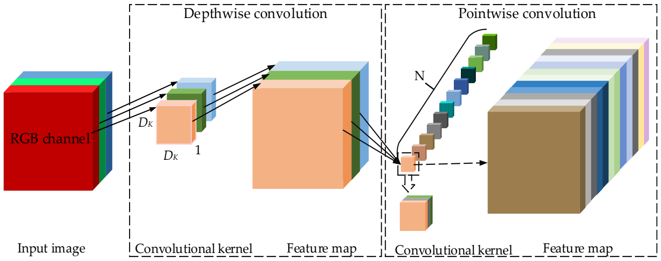

Figure. 4 Depthwise separable convolution graph.

According to reference [35], if we perform the depthwise separable convolution, the parameter quantum of the model are reduced to $1/N+1/D_K^2$ of the original, where N is the number of convolution kernels, D_K is the size of the convolution kernel.

To further reduce the parameter quantum, MobileNetV1 adopts a width factor and a resolution factor, and the range of values is (0, 1). The width factor is used to adjust the number of channels, and the resolution is used to adjust the resolution of the input feature map. The width factor and resolution factor are both 1.

2) *Fast NMS*

Traditional NMS's IoU calculation and sequential iteration suppression result in low computational efficiency. Therefore, the lightweight Mask R-CNN model adopts Fast NMS instead of traditional NMS to solve this problem. Fast NMS parallelizes the IoU calculation process and only calculates the IoU of the bounding box set $B=\{B_i\}_{i=1\ to\ n}$ itself to achieve the NMS suppression function. Sorting the bounding box set B to get the IoU matrix as follows, where B_1 is the highest score frame and B_n is the lowest score frame:

$$X = IoU(B,B) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix}, \qquad x_{ij} = IoU(B_i, B_j) \quad (1)$$

From the symmetry of IoU ($IoU(B_i,B_j)=IoU(B_j,B_i)$), IoU matrix X is a symmetric matrix. And in the process of calculating IoU, it is meaningless to calculate IoU with itself (i.e.i=j), so the IoU matrix X can be upper triangulated. The IoU matrix X with zero diagonal elements and zero lower triangular elements is as follows:

$$X = \begin{pmatrix} 0 & x_{12} & \cdots & x_{1n} \\ 0 & 0 & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \qquad (2)$$

After the upper triangulation process, taking the maximum value of the IoU matrix by column to obtain a one-dimensional tensor $b=[b_1,b_2,\cdots,b_n]$, where b_i is the maximum value of the element on the i-th column. Then Fast NMS sets the threshold to T_h, selects elements in the tensor b, keeps the boxes with elements greater than the threshold in b, and binarizes the tensor b.

3) *Establish a lightweight Mask R-CNN model*

In this paper, the MobileNetV1 network is adopted as the feature extraction layer of the Mask R-CNN model, and Fast NMS is adopted to replace the traditional NMS. A lightweight Mask R-CNN model is designed to meet the actual needs of mobile spruce counting. The network structure of the lightweight Mask R-CNN model is shown in Figure 5.

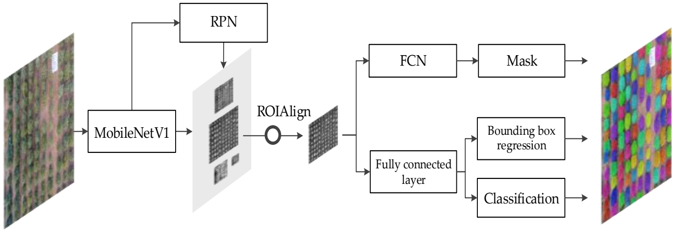

Figure. 5 The Lightweight Mask R-CNN model.

The lightweight Mask R-CNN model mainly includes four parts: feature extraction network, Region Proposal Network, ROIAlign, and output module. First, the feature extraction network extracts the features and outputs the feature map, and then the feature map is input into the RPN for calculations. After going through a 3×3×512 shared convolutional layer with a step size of 1, the output result is input to the two branches. One of the branches generates a candidate region table through convolution operation, and the other branch uses the softmax function to output the probability that the content in the candidate region belongs to the foreground. Finally, Fast NMS removes the candidate regions with lower scores to obtain the final candidate region table. Then input the candidate region table into ROIAlign, which can be transformed into a feature map with a fixed size of 7×7. The feature map then outputs the frame coordinates and classification results through the convolutional layer and the fully connected layer, and the candidate region table outputs the segmentation results through the convolutional layer and the deconvolutional layer. The structure diagram of the frame regression, classification and segmentation network is shown in Figure 6.

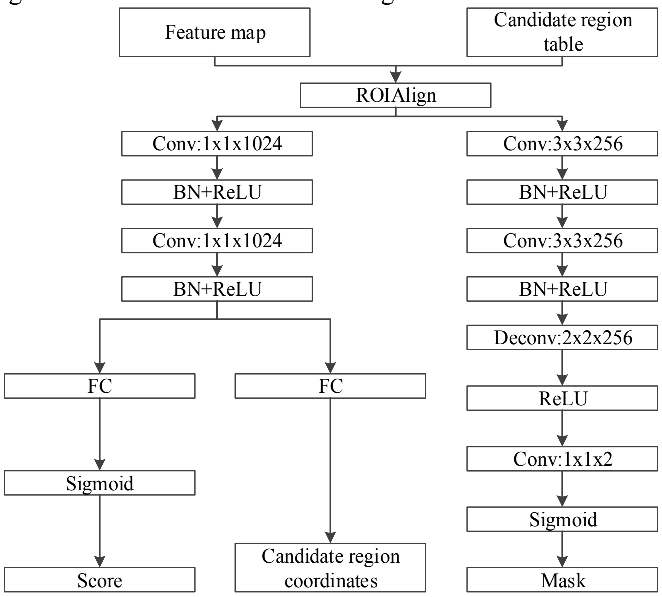

Figure. 6 Bounding box, classification and segmentation network diagram.

In Figure 6, FC represents a fully connected layer, and Deconv represents a deconvolution layer. BN stands for batch normalization, and ReLU stands for ReLU activation function. The network structure of the output module will classify, regress and segment the spruce image at the same time, which is a multi-task network. There will be 5 components in the complete loss function as shown below.

$$L = L_{cls} + L_{box} + L_{mask} + L_{rpncls} + L_{rpnbox} \tag{3}$$

In the formula, L_{cls} represents the classification error, L_{box} represents the frame coordinate regression error, L_{mask} represents the segmentation error, L_{rpncls} represents the RPN layer classification error, and L_{rpnbox} represents the RPN layer regression error.

(1) For the classification error and RPN layer classification error, the cross-entropy function commonly used in classification tasks is selected, and the formula is as follows:

$$L_{cls} = -\frac{1}{m}\sum_{i=1}^{m}\left[p_i \log(p_i) + (1-p_i)\log(1-p_i)\right] \tag{4}$$

In the formula, p represents the probability of correct classification, and m represents the number of candidate regions output by the RPN layer.

(2) For frame regression error and RPN layer regression error, use the smooth L1 function to define, the formula is as follows:

$$L_{box} = \frac{1}{m}\sum_{i=1}^{m}\left[f\left(dx_t, dx_p\right) + f(dy_t, dy_p) + f\left(dw_t, dy_p\right) + f(dh_t, dh_p\right] \tag{5}$$

$$f(a,b) = \begin{cases} 0.5*(a-b)^2/\sigma^2 & |a-b| < 1 \\ |a-b| - 0.5 & |a-b| \geq 1 \end{cases} \tag{6}$$

In the formula, (x_t, y_t, w_t, h_t) represents the marked frame position, and (x_p, y_p, w_p, h_p) represents the frame position predicted by the network.

(3) For the segmentation error, the RPN network will generate m candidate region windows of $a \times a$, the elements of which are probability values in [0, 1]. The error of a single element in the window is defined as follows:

$$L_s\left(y, p\left(y \mid x\right)\right) = -y\ln p(y \mid x) - (1-y)\ln(1-p(y \mid x)) \tag{7}$$

Since the window size is $a \times a$, the complete segmentation is defined as follows:

$$L_{mask} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{a\times a} L_s\left(y_j, p\left(y_j \mid x_j\right)\right) \tag{8}$$

C. Model training and evaluation

1) Model training

The model training hardware platform is E5-2650L V3 8-core CPU, RTX 2080 Ti 11G video memory GPU, 32G running memory, 2TB SSD. Under the Ubuntu16.04 system, using Tensorflow 1.13 and Keras 2.2.4 as the framework, Python language programming is used to build a lightweight Mask R-CNN spruce counting model.

The lightweight Mask R-CNN model training process used transfer learning technology, and initialized the model with model parameters trained on the MS COCO [36] data set to speed up model convergence and reduce the risk of overfitting. We use data enhancement methods to simulate a variety of different shooting situations. For example, the random cropping method simulates the UAV shooting images at different flying heights, the random rotation and random flip methods simulate the UAV's multiple shooting angles, and the random noise method is added to simulate environmental factors such as fog. Finally, we input 1440 images of the training set into the model. All original images will be scaled to 224×224 after entering the lightweight Mask R-CNN model.

According to experience, setting the initial learning rate to 0.0001, the learning rate attenuation to 0.0001, the momentum to 0.9, the batch size to 4, the minimum confidence to 0.5, and the area suggested network anchor frame sizes to 8, 16, 32, 64, and 128. Then use 172 test set images to test the trained lightweight model. The test set also includes spruce images in various situations.

2) Model evaluation

The spruce counting method studied in this paper focuses on the actual counting effect and lightness of the lightweight Mask R-CNN model. Therefore, the lightweight Mask R-CNN is evaluated by five indicators: Mean Counting Accuracy (MCA), Mean Absolute Error (MAE), Mean Square Error (MSE), Average Counting Time (ACT), and Model Size (MS). Among them, MCA refers to the ratio of the number after error removal to the actual number, which is used to characterize the accuracy of the counting algorithm; MAE is the average value of the error, which is used to measure the actual error between the counting result and the actual quantity, and reflects the accuracy of the estimated result as a whole; MSE refers to the expected value of the square of the difference between the model count and the actual quantity, which can be used to evaluate the error between the model count result and the actual quantity, reflecting the stability of the estimated result as a whole; ACT is used to measure the computing speed of the model; MS mainly reflects the complexity of the model and is used to measure whether it is suitable for mobile devices. The calculation process of MCA, MAE, MSE, and ACT is as follows:

$$MCA = \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{|y_i - \hat{y}_i|}{\hat{y}_i} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{11}$$

$$ACT = \frac{t_n}{n} \tag{12}$$

In the above formula, y_i is the true number of spruce in the ith image; $\hat{y}_i$ is the number of spruce in the ith image counted by the model in this paper; n is the number of test images; t_n is the total time spent counting n images.

IV. RESULTS AND DISCUSSION

A. Counting result

Evaluating the performance of the lightweight Mask R-CNN model with MCA, MAE, MSE, ACT, and MS. MCA reaches 95%, MAE is 8.02, MSE is 181.55, ACT is only 1.514 s, and MS is 90Mb.

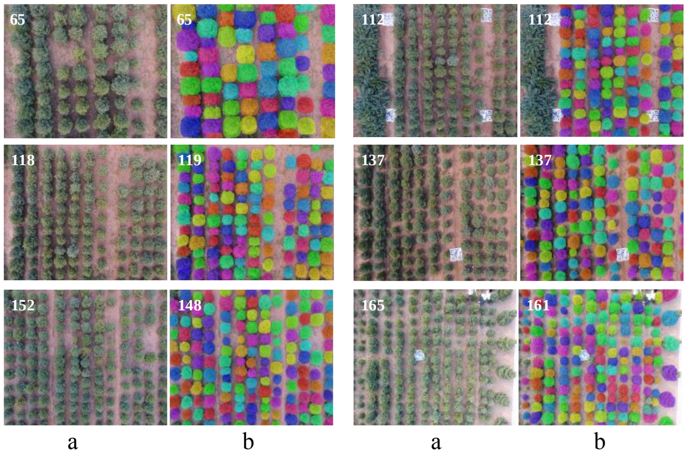

Figure. 7 Lightweight Mask R-CNN counting results. (a) Original image. (b) Lightweight Mask R-CNN.

As shown in Figure 7, the lightweight Mask R-CNN model identifies spruce accurately and has a high recognition accuracy, indicating that the model could identify dense targets. In the case of plant adhesion and a large number of weeds in the test image, the lightweight Mask R-CNN model still effectively avoids interference and segmented different spruce targets. Experiments show that the model is robust to plant adhesion and weed background interference, so it is suitable for spruce counting.

Designing a set of ablation experiments, we can verify the effect of the lightweight Mask R-CNN model we proposed. They are: 1) Original Mask R-CNN model; 2) Only replace the feature extraction network with MobileNetV1 network; 3) Only replace NMS with Fast NMS; 4) Lightweight Mask R-CNN model.

Table. 2 Lightweight effects of the four programs

Programs	**MCA/%**	**MAE**	**MSE**	**ACT/s**	**MS/Mb**
Mask R-CNN	95.6	6.25	102.58	2.39	243
Mask R-CNN+ MobileNetV1	95	8.23	193.9	1.852	90
Mask R-CNN+ Fast NMS	95.6	6.41	118.6	2.073	243
Lightweight Mask R-CNN	95	8.02	181.55	1.514	90

It can be seen from the data in Table 2 that the improvement of the feature extraction network reduces the model size, and the Fast NMS effectively shortens the counting time. The combination of the two improvements completes the lightweight of Mask R-CNN, improves the complex structure of Mask R-CNN, and reduces the amount of calculation. The comparison of MCA, MAE, and MSE shows that the lightweight improvement has little impact on the performance of the network, and it can still maintain better accuracy and lower errors.

B. *Comparative analysis*

To further verify the spruce counting performance of the lightweight Mask R-CNN model, we chose some methods based on deep learning to achieve target counting as a comparison. S.W. Chen from the University of Pennsylvania uses the FCN network to extract candidate regions in the image. A counting algorithm based on a second convolutional network then estimates the number of targets in each region. Finally, use a linear regression model to estimate the total number of targets. This method is called the FCN+Slice counting model below [13]. In addition, based on previous experience, we designed the FCN+Hough circle counting model based on deep learning. J.P. Vasconez et al. used the SSD+MobileNet method to detect and count three kinds of fruits: Hass avocado, lemon, and apples, under different field conditions [37]. This method is called the SSD+MobileNetV1 counting model below. MCA, MAE, MSE, and ACT are used to quantitatively evaluate the counting results of 172 spruce images in the test set.

Table. 3 Counting results of five algorithms

Method	**MCA/%**	**MAE**	**MSE**	**ACT/s**
FCN + Hough circle	85.7	17.88	607.18	3.957
FCN+Slice	89.8	13.04	325.23	3.205
SSD+ MobileNetV1	90.8	12.83	274.62	1.873
Mask R-CNN	95.6	6.25	102.58	2.39
Lightweight Mask R-CNN	95	8.02	181.55	1.514

From the comparison of the counting results of different methods in Table 3, MCA, MAE, MSE, and ACT of the spruce counting method based on the lightweight Mask R-CNN model proposed in this paper are 95%, 8.02, 181.55, and 1.514 s, respectively. Compared with the Mask R-CNN model, the model we proposed has less performance loss while achieving lightweight; compared with the SSD+MobileNetV1 counting model, the FCN+Hough circle counting model and the FCN+Slice counting model, the model we proposed has higher counting efficiency and is more suitable for deployment on mobile terminals.

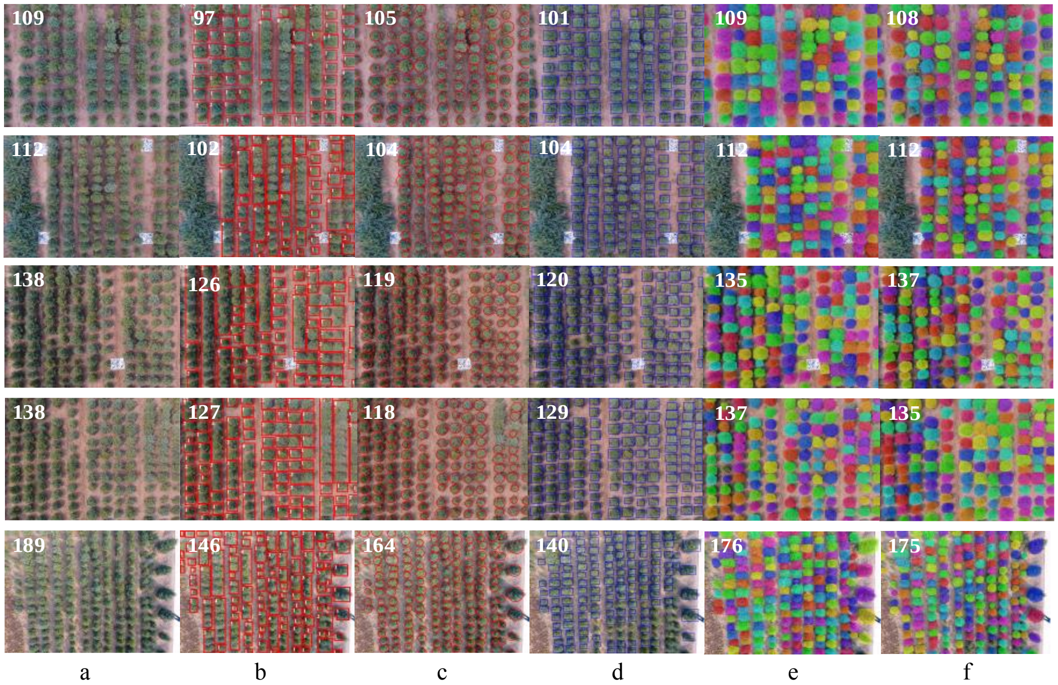

Figure. 8 A comparison of the five models. (a) Original image. (b) The FCN+Slice counting model. (c) The FCN+Hough circle counting model. (d)The SSD+MobileNetV1 counting model. (e) The Mask R-CNN model. (f) The Lightweight Mask R-CNN model.

As shown in Figure 8, the lightweight Mask R-CNN model misses and misidentifies fewer spruce. For images with the uneven size of spruce, the FCN+Hough circle counting model has a large counting error. The FCN+Slice counting model has large errors when counting images with dense spruce and severe adhesion. The SSD+MobileNetV1 counting model has lower counting accuracy when spruce is planted densely in the image. The lightweight Mask R-CNN model completes the accurate segmentation and counting of spruce in the image by effectively solving the problem of spruce adhesion and complex background interference.

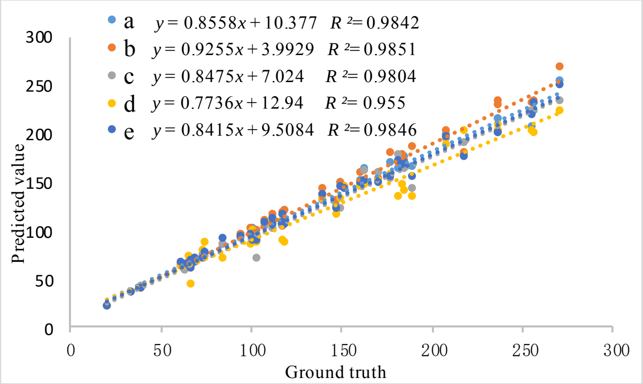

Figure. 9 The linear relationship between the counting results of different methods and the ground truth.
The counting result of (a) the lightweight Mask R-CNN model.(b)the Mask R-CNN model.(c) the FCN+Slice counting model.(d) the FCN+Hough circle counting model.(e) the SSD+MobileNetV1 counting model.

We select 43 original images taken by UAV in the testing set and analyzed the relationship between the counting results of the five counting methods and the ground truth of manual counting by linear correlation analysis. The results are shown in Figure 9. The coefficient of determination R^2 between the counting results of the Mask R-CNN model and the ground truth is 0.9851, and the coefficient of determination between the counting results of the lightweight Mask R-CNN model and the ground truth is 0.9842, both of which are higher than the other three models. The coefficients of determination R^2 for the SSD+MobileNetV1 counting model, the FCN+Slice counting model and the FCN+Hough circle counting model are 0.9846, 0.9804 and 0.955, respectively.

And for the study of spruce counting, the coefficient of determination of the counting results of the lightweight Mask R-CNN model in this paper is only 0.0009 lower than that of the Mask R-CNN model. In the comparative experiment, the coefficients of determination of the five counting methods are ranked from high to low for the Mask R-CNN model, lightweight Mask R-CNN model, the SSD+MobileNetV1 counting model, the FCN+Slice counting model, and the FCN+Hough circle counting model. This ranking is consistent with the MCA ranking. Moreover, it can be seen that the MAE and MSE rankings of the five counting methods are the opposite of the coefficient of determination ranking, which once again confirms the effectiveness of the counting method we proposed. In summary, the combined analysis of MCA, MAE, MSE, and coefficient of determination proves that the lightweight Mask R-CNN model proposed in this paper has a counting accuracy similar to that of the Mask R-CNN model, but is faster and achieves the purpose of optimizing the model.

V. CONCLUSION

In response to the difficulty of counting seedlings in the nursery, taking spruce as the research object, and using UAV as the image acquisition tool, a lightweight Mask R-CNN spruce counting model has been designed. Training the lightweight

Mask R-CNN model with 1440 images of the training set and testing the model with 172 images in the test set. The results show that MCA of this model is 95%, MAE is 8.02, MSE is 181.55, ACT is 1.514 s, and MS is 90Mb. We used the SSD+MobileNetV1 counting model, the FCN+Hough circle counting model, the FCN+Slice counting model, the Mask R-CNN model, and the lightweight Mask R-CNN model to count 172 test set images. In terms of ACT, the lightweight Mask R-CNN model leads other methods by 1.514 s; In terms of MCA, MAE, and MSE, the lightweight Mask R-CNN model is superior to the SSD+MobileNetV1 counting model, the FCN+Hough circle counting model and the FCN+Slice counting model, as well as close to the Mask R-CNN model. And the model size of the lightweight Mask R-CNN model is significantly reduced. The comparative experiment results show that the lightweight Mask R-CNN model has a faster speed, smaller model size, and good accuracy.

After verification, compared with similar research, the lightweight Mask R-CNN model proposed in this paper does not need to be counted in multiple steps, which simplifies the tedious operation process. In addition, the complicated spruce background and individual adhesion can also be overcome, and the characteristics of spruce images can be automatically extracted, which has a high counting accuracy rate. After we made lightweight improvements to the model, we significantly reduced the model size and increased the running speed, which means that the improvement is successful. Next, we will study the application of the lightweight Mask R-CNN model on the mobile terminal, and further research is required to explore the solution of seedling counting in a production unit.

References

[1] M. He, S. Huang, Y. Zhang, M.M. Rahman, “From peasant to farmer: Transformation of forest management in China,” Small-scale Forestry, 2020, vol. 19, no. 2, pp.187-203.

[2] J. Yeom, J. Jung, A. Chang, M. Maeda, J. Landivar, “Automated open cotton boll detection for yield estimation using unmanned aircraft vehicle (UAV) data,” Remote Sensing, 2018, vol. 10, no. 12, pp.1895.

[3] B. Ertugrul, E.B. Muhammed, C.A. Numan, “low-cost UAV framework towards ornamental plant detection and counting in the wild,” ISPRS Journal of Photogrammetry and Remote Sensing, 2020, vol.167, pp.1-11.

[4] P. Chamoso, W. Raveane, V. Parra, A. González, “UAVs applied to the counting and monitoring of animals,” Advances in Intelligent Systems and Computing, 2014, pp.71-80.

[5] W.S. Qureshi, A. Payne, K.B. Walsh, R. Linker, M.N. Dailey, “Machine vision for counting fruit on mango tree canopies,” Precision Agriculture, 2016, vol. 17, no. 3, pp.1-21.

[6] T. Liu, C.M. Sun, L.J. Wang, et al., “In-field wheatear counting based on image processing technology,” Transactions of the Chinese Society for Agricultural Machinery, 2014, vol. 45, no. 2, pp.282-290.

[7] J.B. Scott, D.H. Gent, F.S. Hay, S.J. Pethybridge, “Estimation of pyrethrum flower number using digital imagery,” HortTechnology, 2015, vol. 25, no. 5, pp.617-624.

[8] L. Fu, E. Tola, A. Al-mallahi, R. Li, Y.J. Cui, “A novel image processing algorithm to separate linearly clustered kiwifruits,” Biosystems Engineering, 2019, vol. 183, pp.184-195.

[9] J.D. Lv, D.A. Zhao, W. Ji, S.H. Ding, “Recognition of apple fruit in natural environment,” Optik, 2016, vol. 127, pp.1354-1362.

[10] Z.S. Pothen, S. Nuske, “Texture-based fruit detection via images using the smooth patterns on the fruit,” In Proc. 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16-21 May, 2016, pp.5171-5176.

[11] W.S. Qureshi, S. Satoh, M.N. Dailey, M. Ekpanyapong, “Dense segmentation of textured fruits in video sequences,” In Proc. 9th International IEEE Conference on Computer Vision Theory & Applications, Lisbon, Portugal, 5 - 8 January, 2014, pp.441-447.

[12] R. Linker, O. Cohen, A. Naor, “Determination of the number of green apples in RGB images recorded in orchards,” Computers and Electronics in Agriculture, 2012, vol. 81, no. 1, pp.45-57.

[13] S.W. Chen, S.S. Shivakumar, S. Dcunha, et al., “Counting Apples and Oranges With Deep Learning: A Data-Driven Approach,” IEEE Robotics & Automation Letters, 2017, vol. 2, no. 2, pp.781–788.

[14] J. Long, E. Shelhamer, T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, vol. 39, no. 4, pp.640-651.

[15] W. Li, H. Fu, Y. Le, C. Arthur, “Deep learning based oil palm tree detection and counting for high-resolution remote sensing images,” Remote Sensing, 2016, vol. 9, no. 1, 22.

[16] F.J. Chen, X.Y. Zhu, W.J. Zhou, M.M. Gu, Y.D. Zhao, “Spruce counting method based on improved YOLOv3 model in UVA images,” Transactions of the Chinese Society of Agricultural Engineering, 2020, vol. 36, no. 22, pp.22-30.

[17] M, Rahnemoonfar, C. Sheppard, “Deep Count: Fruit Counting Based on Deep Simulated Learning,” Sensors, 2017, vol. 17, pp.905.

[18] K. He, G. Gkioxari, P. Dollar, “Mask R-CNN,” In Proc. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017, pp.2980–2988.

[19] M. Machefer, “Mask R-CNN Refitting Strategy for Plant Counting and Sizing in UAV Imagery,” Remote Sensing, 2020, vol. 12, no. 18, pp.3015-.

[20] Y. Tian, G. Yang, Z. Wang, E. Li, Z. Liang, “Instance segmentation of apple flowers using the improved mask R-CNN model,” Biosystems Engineering, 2020, vol. 193, pp.264-278.

[21] N.E. Ocer, G.J. Kaplan, F. Erdem, D.K. Matci, U. Avdan, “Tree extraction from multi-scale UAV images using Mask R-CNN with FPN,” Remote Sensing Letters, 2020, vol. 11, no. 9, pp.847-856.

[22] B. Xu, W. Wang, G. Falzon, P. Kwan, C. Li, "Livestock classification and counting in quadcopter aerial images using Mask R-CNN," International Journal of Remote Sensing, 2020, no. 7, pp.1-22.

[23] H. Jiang, C. Zhang, Z. Zhang, et al., "Detection Method of Corn Weed Based on Mask R-CNN," Transactions of the Chinese Society for Agricultural Machinery, 2020, vol. 51, no. 6, pp.220-228, 247.

[24] J. Deng, Z. Zhong, H. Huang, et al., "Lightweight Semantic Segmentation Network for Real-Time Weed Mapping Using Unmanned Aerial Vehicles," Appied Science, 2020, vol. 10, no. 20, pp.7132.

[25] A.G. Howard, M. Zhu, B. Chen, et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861v1, 2017.

[26] A. Neubeck, A.; L. Gool, "Efficient Non-Maximum Suppression," In Proc. 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China, 20-24 August 2006, pp.850–855.

[27] D. Bolya, C. Zhou, F.Y. Xiao, Y.J. Lee, "YOLACT: Real-time instance segmentation," In Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October-2 November 2019, pp. 9156-9165.

[28] G. Pradeep Kumar, B. Sridevi, "Simulation of Efficient Cooperative UAVs using Modified PSO Algorithm," WSEAS Transactions on Information Science and Applications, 2019, vol. 16, Art. #11, pp. 94-99.

[29] Lucjan Setlak, Rafal Kowalik, "Control Model of a Small Micro-class UAV Object Taking Into Account the Impact of Strong Wind," WSEAS Transactions on Systems and Control, 2019, vol. 14, Art. #50, pp. 411-418.

[30] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," In Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016, pp. 770–778.

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, "Aggregated residual transformations for deep neural networks," In Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 1492–1500.

[32] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In Proc. 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.

[33] Y. Liu, Q. Feng, S.Z. Wang, "Plant disease identification method based on lightweight CNN and mobile application," Transactions of the Chinese Society of Agricultural Engineering, 2019, vol. 35, no. 17, pp.194-204.

[34] X. Wu, Z.Y. Qi, L.J. Wang, J.J. Yang, X. Xia, "Apple detection method based on Light-YOLOv3 convolutional neural network," Transactions of the Chinese Society for Agricultural Machinery, 2020, vol. 51, no. 8, pp.17-25.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," In Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 1251–1258.

[36] T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, "Microsoft COCO: Common Objects in Context," In Proc. 2014 European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014, Springer, Cham, Switzerland, 2014, pp. 740–755.

[37] J.P. Vasconez, J. Delpiano, S. Vougioukas, F. Auat Cheein, "Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation – ScienceDirect," Computers and Electronics in Agriculture, 2020, vol. 173.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Wenjing Zhou and Xueyan Zhu contributed to methods design, data analysis and drafted the manuscript. Fengjun Chen conceived the technical solution. All authors have read and agreed to the published version of the manuscript.

Sources of funding for research presented in a scientific article or scientific article itself

National Key Research and Development Program of China (2019YFD1002401)

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)