CountShoots: Automatic detection and counting of slash pine new shoots using UAV imagery

Xia Hao[1], Yue Cao[1], Zhaoxu Zhang[1], Federico Tomasetto[3], Weiqi Yan[4], Cong Xu[5], Qifu Luan[2], Yanjie Li[2*]

1. College of Information Science and Engineering, Shandong Agricultural University, N0. 61, Daizong Road, Taian, 271018, Shandong Province, China

2. Research Institute of Subtropical Forestry, Chinese Academy of Forestry, No. 73, Daqiao Road, Fuyang, Hangzhou 311400, Zhejiang Province, China

3. AgResearch Ltd., Christchurch 8140, New Zealand

4. Department of Computer Science, Auckland University of Technology, Auckland 1010, New Zealand

5. School of Forestry, University of Canterbury, Private Bag 4800, 8041 Christchurch, New Zealand

**Corresponding author: Yanjie Li: aj7105@gmail.com*

Abstract

The density of new shoots on pine trees is an important indicator of their growth and photosynthetic capacity. However, traditional methods to monitor new shoot density rely on manual and destructive measurements, which are labor-intensive and have led to fewer studies on new shoot density. Therefore, in this study, we present user-friendly software called

CountShoots, which extracts new shoot density in an easy and convenient way using unmanned aerial vehicles (UAV) based on the YOLOX and SPSC-net models.

This software mainly consists of two steps. Firstly, we deployed a modified YOLOX model to identify the tree species and location from complex RGB background images, which yielded a high recognition accuracy of 99.15% and 95.47%. These results showed that our model produced higher detection accuracy compared to YOLOv5, Efficientnet, and Faster-RCNN models. Secondly, we constructed a Slash Pine Shoot Counting Network (SPSC-net). This methodology is based on the CCTrans network, which outperformed DM-Count, CSR-net, and MCNN models, with the lowest MSE and MAE results among other models (i.e. 2.18 and 1.47, respectively).

To our best knowledge, our work is the first research contribution to identify tree crowns and count new shoots automatically in Slash Pine. Our research outcome provides a highly efficient and rapid user-interactive pine tree new shoot detection and counting system for tree breeding and genetic use purposes.

Keywords: deep learning; Object Detection; tree crown; new shoots; software

1 Introduction

In southern China, the slash pine (*Pinus elliottii*) is an exotic pine species, which has become one of the most domesticated tree species mainly for timber and resin production (Ding, et al. 2022,Lai, et al. 2020). One key reason for the domestication of slash pine is its genetic improvement (Pagliarini et al., 2020), which ensures sustainable wood timber and resin production, and enables the selection of productive breeds through constant monitoring of growth traits. One of the most important traits, new shoot density, i.e. the number of new shoots per tree, should be one of the main focuses in the tree breeding strategy. The new shoots of pine trees play an important role in tree growth, nutrition accumulation, and re-translocation (Fife and Nambiar 1982). Shoot density is closely associated with nutrient absorption and re-translocation, tree growth, crown size and crown photosynthetic capacity, which could be an important indicator for the selection of timber and resin yield of productive tree (Kellomäki and

Strandman 1995,Fife and Nambiar 1984,Whitmore and Zahner 1966).

A standard way to measure the new shoot density of pine trees is to count in each tree, which relies on human observation for the tree breeding process that is inefficient and time-consuming (Li, et al. 2020). In addition, due to the tree height and crown density, it is difficult to count the new shoot in an efficient way. Therefore, most studies so far have focused on the length of new shoots (Zweifel, et al. 2020) and less on new shoot density. To the best of our knowledge, the tree new shoot density and the absolute number have been less studied (Stadler, et al. 2005,Rosati, et al. 2018), and these studies primarily relied on manual counting methods, lacking an efficient and automated approach for quantifying shoot density and absolute numbers. It is therefore needed to develop automated new shoot detection techniques based on deep learning methodology to meet the modern requirement of high-throughput and efficient measurement of tree traits.

With the development of inexpensive and highly efficient unmanned aerial vehicles (UAV) camera platforms, in-filed RGB image-based target detection has successfully emerged as a powerful and reliable solution to

substitute laborious traditional manual measurement in agriculture and forestry studies (Dalla Corte, et al. 2020,Picos, et al. 2020,López-Granados, et al. 2019). UAV-based high definition RGB imagery holds the advantage in obtaining high resolution target images in a high throughput way and produces a relatively high accuracy coupled with the deep learning methodology (Xie and Yang 2020). The main process involves using a computer-based system to simulate human visual functions in order to detect relevant feature information from RGB images, ultimately achieving high-accuracy detection of plant targets. (Pound, et al. 2017). It has been successfully applied to the detection of the wheat head (Sun, et al. 2022), ornamental plant (Bayraktar, et al. 2020), wheat yellow rust (Su, et al. 2018) and Gramineae weed detection (Barrero and Perdomo 2018).

The rapid development of computer software and hardware has dramatically improved the performance of computer vision technology for the application of various plant traits. Machine learning (ML) and deep learning (DL) methods are the two mainstream methods for plant trait counting. ML, such as Support Vector Machines (SVM) (Evgeniou and Pontil 1999) and Random Forests (RF) (Breiman 2001), has been

harnessed to build regression or classification models for plant target detection and counting by extracting features such as color and texture from images (Singh, et al. 2018,van Dijk, et al. 2021). However, ML requires human-defined features, and as the data increases, the performance is saturated, which is not suitable for the growing demands of large data processing (Chlingaryan, et al. 2018,Li, et al. 2022).

DL is a branch of machine learning that utilizes neural networks to process large amounts of data, reduce prediction errors, and automatically extract features (LeCun, et al. 2015,Bengio, et al. 2013). It has been widely applied in plant trait detection to meet the requirement of accurate plant trait counting (Kamilaris and Prenafeta-Boldú 2018).

The mainstream counting method of deep learning uses convolutional networks to regress density maps, which often has difficulties to capture global features for global context modeling and usually requires the introduction of additional attention mechanisms, with the model structure which is gradually complicated (Chen et al., 2019). Vision Transformer (ViT; Dosovitskiy, et al. 2020), with its powerful global context modeling capabilities, shows powerful processing capabilities in dense prediction

tasks such as object detection and segmentation. Liang, et al. (2022) proposed TransCrowd, which reformulates the problem of weakly supervised population counting from the perspective of ViT-based sequence counting, is also the first model for weakly supervised population counting based on Transformer. Instead of generating predicted density maps, it trains directly from images to counts in a weakly supervised manner, based on the number of people in the images. Inspired by classification, BCCT (Sun et al, 2021) added a context token in the input sequence to enable better information exchange between the transformer in the model and image patches. Compared with the work of TransCrowd, the overall network processing method by Sun et al. (Sun et al. 2021) is simplified and better performance is achieved. Another counting model, CCTrans (Tian, et al. 2021), also has been proposed, with its crowd counting model using Twins as the backbone network (Chu, et al. 2021). The Feature Pyramid Fusion Module (FPA) is used in the CCTrans model to supplement detailed information with low-level features, resulting in crowd features that are rich in semantics, detailed information, and global features. Finally, a multi-scale dilated convolution module (MDC) is

designed as the regression head to deal with the global features captured by the Transformer, which is very helpful in regressing a more precise and accurate crowd density map. This method optimizes the loss function of the mainstream with strong and weak supervised forms to improve the generalization performance and robustness of the model.

Therefore, in this paper we propose to construct a Slash Pine Shoot Counting Network (SPSC-net) model based on CCTrans to count the new shoots of slash pine. We introduced the feature pyramid module to achieve accurate counting of slash pine new shoot. In summary, the aims of this paper are to 1) Construct a target extraction network to accurately detect and extract the individual slash pine trees from the complex background; 2) Create a multi-scale slash pine new shoot counting network SPSC-net, which is based on CCTrans and introduces non-equilibrium transmission and perspective guidance to improve the loss function in the fully supervised process. As a result, it efficiently and accurately completes the counting task in multi-scale wetland pine images; 3) create a highly efficient user-interactive pine tree new shoot detection and counting system for the convenient of tree breeding and genetic use purpose.

2 Methods and materials

In this paper, we built a framework for automatically counting new shoots on UAV images. The UAV images containing one or more plants were used as input. The framework (Figure 1) was developed in two stages. The first step is the extraction of a single plant, which builds a target extraction network based on a manually labeled data set to realize the detection and segmentation of a single slash pine, in order to obtain a single target and discard complex backgrounds. The second step is to count the new shoots. This method is needed to generate a density map based on the center coordinate points of the manually labeled tipping as an intermediate representation to supervise the learning of the counting model. The trained counting model can directly predict the number of shoots. More details are described in the following sections.

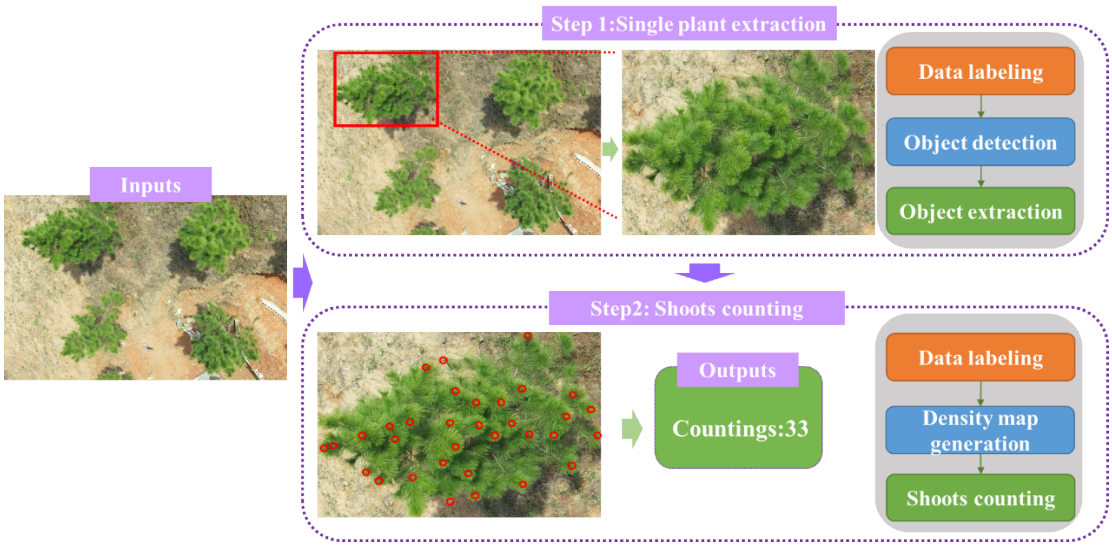


Figure 1 The overall research roadmap

2.1 Site location and images collection

The site for a breeding population of Slash pine is located at Matou national forest farm in Xuancheng, Anhui, China (30◦45′N, 118◦29′E), covering an area of approximately 49.4 acres. The information on this site was described by Song, et al. (2022). In brief, this site is located in a subtropical climate, with precipitation and average yearly temperature of 1520 mm and 15.7 ◦C respectively. The images of each slash pine tree were collected in March and June 2022 respectively using DJI Phantom 4 RTK UAV with a 4864×3648 resolution RGB camera (DJI, Shenzhen, Guangdong, China). To obtain images of new shoots at different growth stages, two flights per month were conducted in both March and June, with an interval of about

10 days between each flight. To avoid the severe sunlight affecting the image quality, cloudy days were chosen for each flight. The flight altitude was set at 20 m at 90 degrees, with side and front image overlap percentages of 80% and 85%, respectively. Each image has a resolution of 5472×3648 dpi. After removing images with severe blur and distortion, a total of 1860 images were applied for data analysis. A subset of images of slash pine is shown in Figure 2.

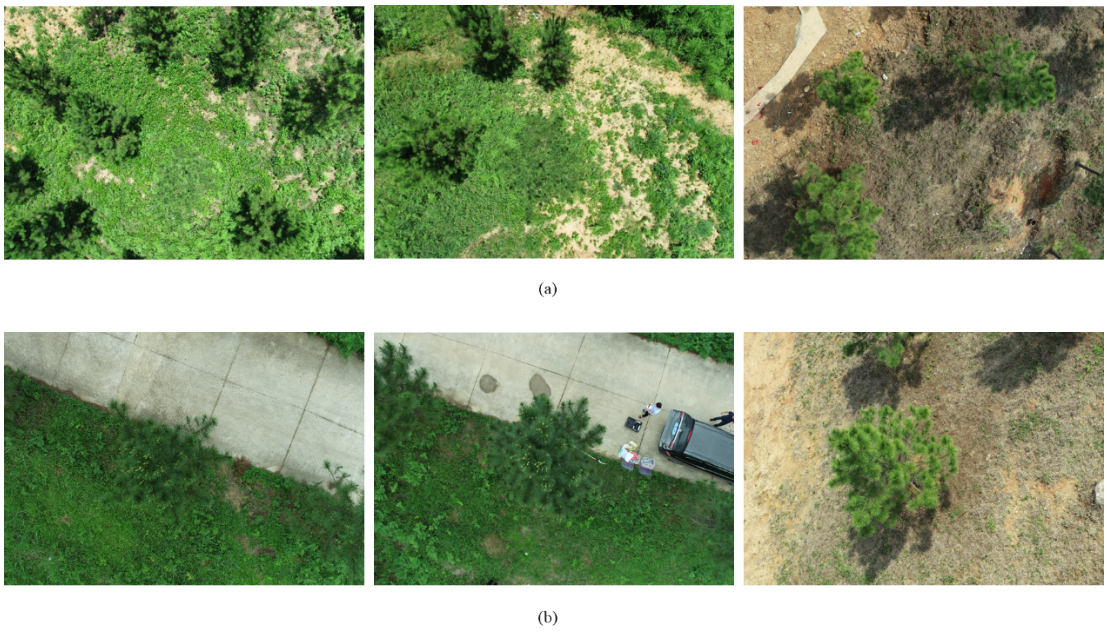


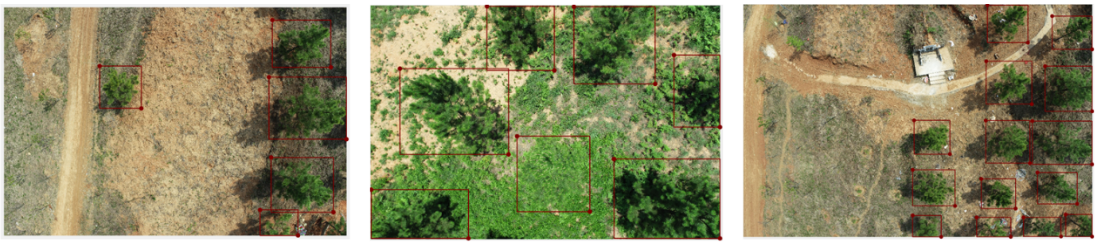
Figure 2 Examples of original UAV images. (a) with multiple slash pine trees; (b) with single slash pine trees.

In both experiments (i.e., single plant extraction and new shoots counting), the original dataset was divided into training, validation, and test sets in a ratio of 7:1.5:1.5. In the object detection task, the number of original

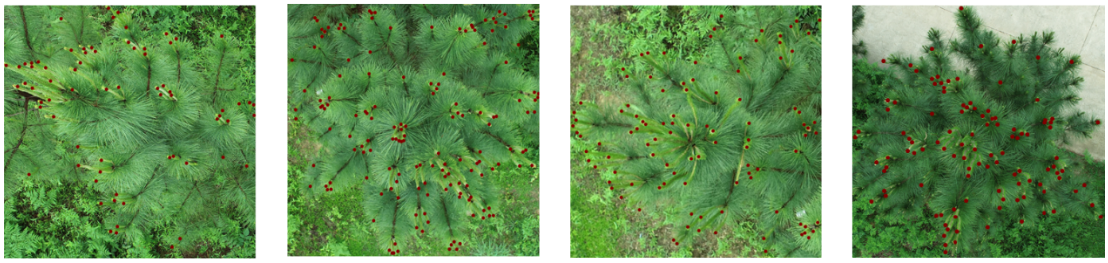
images in the training, validation, and test sets were 1302, 279, and 279, respectively. In the counting task, the numbers were 219, 47 and 47 for the training, validation, and test sets, respectively.

2.2 Data labeling

According to the two processes of shoot counting, two types of labeling were created, including every single tree and the new shoots in each tree (Figure 3 a-b). Each label in Figure 3(a) records the position of the circumscribed rectangle, which is annotated with LabelImg (Tzutalin 2015). The section involved labeling a total of 1860 images and 8488 bounding boxes. The labels in Figure 3(b) show the point position of each new shoot along with the number of the whole plant, which is annotated by Matlab. We tackled the issue of counting by labeling 313 images with 36,758 points.



(a)



(b)

Figure 3 Example of data labelling. (a) Labeling of individual plant extracts; (b)Labeling of new shoots counting

2.3 Density map generation

The network can deal with complicated scenes and high-density tree crown overlapping by using the density map as the intermediate representation to supervise the learning of deep learning models. Therefore, a counting network based on a density map was designed. In order to map the point coordinates obtained by labeling the new shoots of slash pine into a true value density map, a geometrically adaptive kernel method (Yin, et al. 2010) was used. In Figure 4, the dark blue part indicate the background,

and the highlighted warm color area represents the densely distributed new shoots.

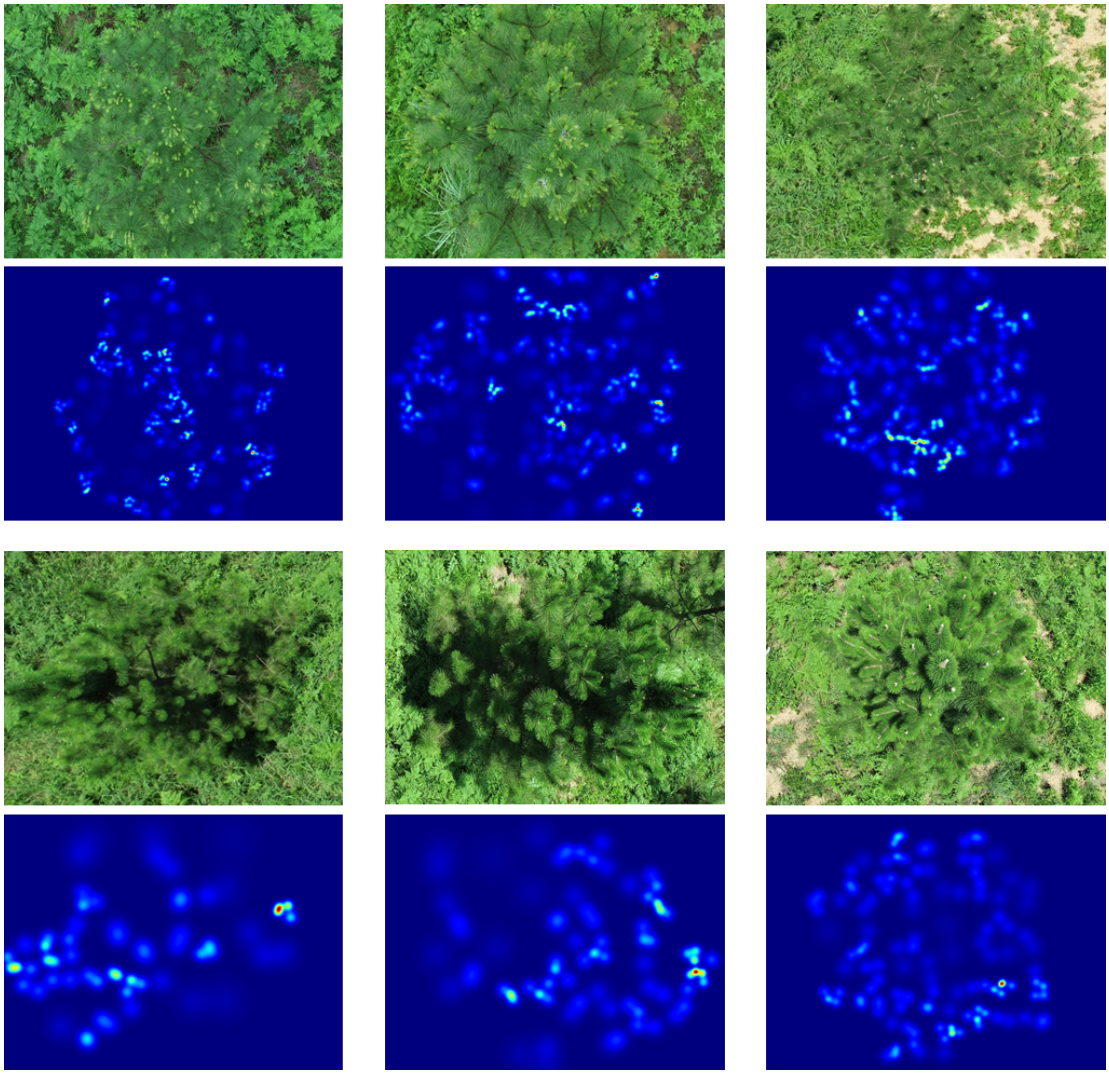


Figure 4 Samples of ground-truth density map for counting labels

2.4 Single plant extraction network

We introduced a YOLOX algorithm (Ge, et al. 2021) as the basic network structure to obtain the detection and extraction of single slash pine. This algorithm was adjusted according to the requirements of detection tasks,

including backbone, neck, and head network respectively. The backbone network of YOLOX followed the Darknet network of YOLOv3 (Redmon and Farhadi 2018). Firstly, the model downsampled the input data through the focus structure to reduce the computational overhead required for model inference. The Neck network of the plant extraction model was based on the image Feature Pyramid Network (FPN) (Zhu, et al. 2020). This detects slash pine trees at different scales and accurately identifies the slash pine trees with overlapping. A neck network obtained the feature map of the extraction target from three different scales for regression analysis. This added an IoU branch into the regression branch to complete the extraction task of single slash pine. During the training process of the model, we set the epoch to 200 and the learning rate to 0.01.

2.5 New shoots counting model

At present, the mainstream method for counting objects is convolutional neural network (CNN). However, due to the limited receptive field, it is difficult to capture global features for context modeling that is often necessary to introduce additional attention mechanisms (Vaswani et al.

2007), which also complicates the model structure (Liang et al., 2023). Therefore, we constructed a Slash Pine Shoot Counting Network (SPSC-net). This methodology, which is based on the CCTrans network (Tian, et al. 2021), generates a generalized loss function to learn the density map for shoot counting and location through the non-equilibrium optimal transfer. The main process is as follows:

For a fully supervised process, CCTrans is based on the design of the loss function proposed by Wang, et al. (2020), which is composed of a weighted sum of counting loss, optimal transport (OT) loss, and total variation (TV) loss. As the TV loss uses the original head annotation from ground truth, which cannot be used to construct the segmentation target, the mean square error (L_2) is used as a substitute for $\mathcal{L}_{TV}$. The original function treats the density map and the point lattice map as probability distributions and uses balanced OT to match the shapes of these two distributions. The point lattice map refers to a discrete grid or lattice representation, where each grid cell represents a specific location or region in the image space. Here, the point lattice map serves as a reference or target for the density map, enabling the comparison and alignment of their distributions using OT. For

the predicted density map D and the ground truth D', the total loss is calculated as follows:

$$\mathcal{L}_d = L_1(P,G) + \lambda_1 \mathcal{L}_{OT} + \lambda_2 L_2(D,D') \quad (1)$$

Where P and G denote the number of new shoots of D and D', and λ_1 and λ_2 are the loss coefficients. The loss function L_1, which corresponds to the smoothed L_1, is used to improve the robustness of the network. In addition, the smoothed annotation map, which refers to the density map after a smoothing process, is utilized to enhance the network's robustness. Furthermore, the mean square error (*i.e.* L_2) is used to adjust the gap between the prediction and the smoothed annotation map. $\mathcal{L}_{OT}$ is the OT loss, which will be discussed in the following paragraph. One of the main shortcomings of the CCTrans network is that normalization of the density map predictions and point lattice maps for computing balanced OT removes the actual counts in both maps. This requires additional counting loss to ensure accurate prediction of counts, which is the sum of the density map. The balanced OT loss is used to align the predicted density map with the ground truth density map, while the counting loss is used to encourage accurate count predictions. The counting loss provides poor monitoring

because its gradient adds the same outlier to all pixels, regardless of their relevance to the crowd count. This means that the loss may prioritize minimizing errors in unimportant regions, while neglecting errors in regions that are crucial for accurate counting. We proposed to consider non-equilibrium OT in our approach because it ensures that any discrepancies between the predicted and Ground Truth (GT) annotations are penalized. This is achieved by our pixels and point loss terms, which directly monitor incorrect predictions. Therefore, any imbalances in the process are accounted for and minimized during optimization. Formally, suppose the predicted density plot is:

$$\mathcal{A} = \{(a_i,\ x_i)\}_{i=1}^{n} \tag{2}$$

Here, $a_i$represents the predicted pixel density, and $x_i \epsilon \mathbb{R}^2$ represents the number of pixels. We used $a = [a_i]_i$ which represents the predict density map. The ground reality dot plot, represents the actual data points, can be shown as:

$$\mathcal{B} = \{(b_j,\ y_j)\}_{j=1}^{m} \tag{3}$$

In a dot plot, each data point is represented by a dot on a two-dimensional coordinate system. Where y_j is the position of the j_{th} marking, m is the

number of marking points, and b_j is the number of target represented by the marking. This plot provides a visual representation of the spatial distribution and quantity of the targets in the dataset. In this study, we assumed: $b = [b_j]_j = 1_m$, Our loss function was based on the optimal transmission cost of the imbalance of entropy regularization, as follows:

$$\mathcal{L}_C^{\mathcal{T}}(\mathcal{A},\mathcal{B}) = \min_{P \in \mathbb{R}_+^{n \times m}} \langle C,P \rangle - \varepsilon H(P) + \tau D_1(P1_m \,|\, a) + \tau D_2(P^T 1_n \,|\, b) \quad (4)$$

where $C \in \mathbb{R}_+^{n \times m}$is the transportation cost matrix; P is the transmission matrix, which will get the best shipping costs from $\mathcal{A}$ to $\mathcal{B}$ at each location by minimizing the treatment; $\langle C,P \rangle$ is the transmission loss, which pushes the predicted density towards the label during training. This loss term characterizes the difference between two distributions, similar to the OT loss in the original loss function. $H(P) = \sum_{ij} P_{ij} \log P_{ij}$ is the entropy regularization term, which encourages balanced or uniform distribution in the transportation scheme and is related to the unbalanced OT loss. When the value of ε is larger, the predicted density map becomes less compact. The TV loss term $\tau D_1(P1_m \,|\, a) + \tau D_2(P^T 1_n \,|\, b)$ includes the product terms of matrices D1 and D2, as well as the matrix P normalized by rows and columns. D_1 constrains the row sums of the

transportation matrix P to match the GT point-annotated density map, while D_2 provides auxiliary constraints to align the column sums of P with the reconstruction of GT point annotations. This complements the counting loss and further improves the accuracy of density map prediction. In which, $\hat{a}$=$P1_m$is the intermediate density map representing the GT annotation construction; $\hat{b}$=$P^T1_n$is the reconstruction of the GT point annotation. Compared to fixed transport with fixed parameters, adaptive transport adjusts automatically during the transportation process based on the network model, enabling more accurate prediction of density maps. During the experiment, ε was set to 0.05 and τ was set to 0.5.

In view of the problem that TV loss is prone to overfitting, we used squared L2-norm ($\|.\|_2$) for the pixel-wise term and L1-norm ($\|.\|_1$) for the point-wise term:

$$D_1(P1_m \,|\, a) = \|P1_m - a\|_2^2 \tag{5}$$

$$D_2\left(P^T1_n \,\middle|\, b\right) = \left\|P^T1_n - b\right\|_1 \tag{6}$$

In addition, the original loss used the standard square Euclidean distance as the transportation cost, equally considering all distances. Considering the perspective effect in the slash pine shoot image, the new shoots close

to the drone will appear sparser, and the new shoots further away from the UAV will appear dense in the image. In order to better represent the density of the new shoot at a distance, the transportation cost of these areas should be higher. Therefore, this study proposed a perspective-based transfer cost matrix to address the scale variation in the image. The cost function is defined as:

$$C_{ij} = exp\left(\frac{1}{\eta(x_i, y_j)}\|x_i - y_j\|_2\right) \quad (7)$$

where $\eta(x_i, y_j)$ is the adaptive perspective coefficient, In order to adapt to the variations of various factors in different shooting scenes, the value of $\eta(x_i, y_j)$ is determined as the average of the normalized heights of the two target pixels in their respective images, that is, $\frac{1}{2}(h_{x_i} + h_{y_i})$. In addition, The Euclidean distance, represented by the cost function $L_{ij} = ||x_i - y_i||_2$, is used to measure the distance cost between two points. This formula is essentially exponentializing the original Euclidean distance divided by an adaptive perspective factor, which amplifies the transmission matrix above the image (often the wetland pine canopy) and further improves the accuracy of shoot counting.

2.6 Model evaluation

2.6.1 Single tree detection model

Recall, precision and recognition accuracy (AP) were employed as the evaluation indexes of the single tree extraction model. Recall is the proportion of all positive samples in the test set that are correctly identified as positive samples. Accuracy refers to the proportion of actual positive samples in the target detected as positive samples, with recall and precision as the horizontal and vertical coordinates to form a Precision-Recall (P-R) curve. The area under the entire P-R curve is AP: the higher AP is, the better its detection performance. The recall, precision, and AP equations are as:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$\mathrm{AP} = \int_0^1 p(r)dr \tag{10}$$

where TP represents the number of correctly detected detection, FP shows the number of detections detected by false detection, FN indicates the number of detection distinguished by leakage, and $p(r)$ reveals the P-R curve respectively.

2.6.2 New shoots detection and counting model

The mean squared error (MSE) and mean absolute error (MAE) were accommodated as evaluation indicators for the new shoots detection and counting model. MSE is generally used to detect the deviation between the predicted value and the true value of the model. MAE is the average of the absolute error, reflecting the actual error of the predicted value, calculated as follows:

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}|P_i - G_i| \tag{11}$$

$$\text{MAE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}|P_i - G_i|^2} \tag{12}$$

where n is the number of test images, P_i, and G_i indicates the predicted and actual number of new shoots in the i_{th} image respectively.

In order to prove the effectiveness of the SPSC-net used in this study it was compared with other classical counting algorithms including DM-Count (Özleyen and Aptoula 2021), CSR-net (Li, et al. 2018), and MCNN. The evaluation algorithm was set according to the environment and parameters of SPSC-net, the batch size was set to 16, and the initial learning rate was set to 10^-5.

All of the analysis in this study was conducted using PyCharm 2021 and

the PyTorch 1.11.0 deep learning framework. The computer ran 128G of memory, the CPU model is Inter(R) Core(TM) i7-10700k, equipped with NVIDIA GeForce RTX 3060, the operating system was Windows 10, and the CUDA version is 11.3.109.

3 Results

3.1 Slash pine tree detection

As shown in Table 1, the threshold, set to 0.5, is used as a criterion to determine the tree detection results. It serves as a threshold for the predicted probabilities, where predictions with probabilities above 0.5 are classified as trees. By setting the threshold to 0.5, the tree detection precision and recall of YOLOv5, Efficientnet, and YOLOX are approximately 90%. When the threshold increased to 0.75, precision, recall, and AP of YOLOX was the best in all models, with APs of 99.29% and 96.08% respectively. In comparison, YOLOv5 performs slightly worse at higher thresholds, but still outperforms Efficientnet and Faster-RCNN. Faster-RCNN performs the worst among all models, with recall, precision, and AP values all below 30% at the 0.75 threshold. For the purpose of

objectively evaluating the detection performance of various models under the conditions of overlapping targets, blurriness, and complex background, a manual counting of the specific detection situation for 26 test images was conducted. Table 2 shows statistical information for 26 test images, including the total object count, the number of correctly detected objects, false detections, and missed detections. YOLOX and YOLOv5 showed a relatively low false detection rate. EfficientNet had the lowest number of missed targets, but the false detection rate was slightly higher, mainly due to some targets being detected repeatedly. Overall, YOLOX exhibited the best performance most correctly detected objects, while Faster-RCNN performed the worst in this regard.

Table 1 The model performance of different tree detection models

Model	Threhold=0.5			Threhold=0.75		
	Recall	Precision	AP	Recall	Precision	AP
YOLOX	99.01%	90.21%	99.29%	96.21%	87.66%	96.08%
YOLOv5	96.84%	90.09%	98.16%	89.17%	82.96%	87.42%
Efficientnet	97.75%	87.43%	97.80%	92.81%	83.01%	91.20%
Faster-RCNN	71.20%	27.78%	57.53%	29.59%	11.55%	13.88%

Some results of the four models on the detection of slash pine trees canopy have been displayed in Figure 5. It shows that YOLOv5, Efficientnet and

Faster-RCNN have issues with missed detections (yellow marked area). Moreover, in dense tree crown areas, Faster-RCNN also suffers from significant problems with repeated detections (indicated by the blue marked areas). However, YOLOX shows the best detection results in complex backgrounds and overlapping targets.

Table 2 Partial object detection results of different models

Image ID	Total Boxes	YOLOX			YOLOv5			Efficientnet			Faster-RCNN		
		Correct Detection	False Detection	Missed Detection	Correct Detection	False Detection	Missed Detection	Correct Detection	False Detection	Missed Detection	Correct Detection	False Detection	Missed Detection
1	13	13	0	0	13	0	0	13	0	0	8	3	5
2	15	15	0	0	15	0	0	15	0	0	9	12	6
3	10	10	0	0	10	2	0	10	3	0	7	2	3
4	5	5	0	0	5	1	0	5	4	0	4	4	1
5	5	5	0	0	5	1	0	5	5	0	4	1	1
6	2	2	0	0	1	0	1	2	3	0	2	4	0
7	6	6	1	0	2	0	4	6	1	0	5	2	1
8	7	7	1	1	3	0	4	7	1	0	5	2	2
9	8	6	0	2	3	0	5	8	0	0	4	2	4
10	5	5	0	0	5	2	0	5	4	0	4	8	1
11	8	8	1	0	8	0	0	8	0	0	6	11	2
12	9	9	0	0	9	1	0	9	1	0	5	5	4
13	14	14	0	0	14	1	0	14	4	0	6	3	8
14	16	16	0	0	16	0	0	16	2	0	12	6	4
15	12	12	0	0	12	0	0	12	2	0	6	4	6
16	17	17	0	0	17	0	0	16	1	0	10	5	7
17	11	11	0	0	11	1	0	11	1	0	5	4	6
18	6	6	0	0	6	0	0	6	0	0	5	7	1
19	7	7	3	0	7	2	0	7	5	0	5	4	2
20	13	13	1	0	13	0	0	13	0	0	4	4	9
21	24	24	0	0	24	0	0	24	1	0	10	9	14
22	36	35	1	1	31	0	5	34	4	2	5	8	31
23	10	9	1	0	8	0	2	10	2	0	6	7	4
24	8	8	1	0	8	0	0	8	1	0	3	4	5
25	19	16	1	3	15	0	4	19	2	0	4	5	15
26	17	17	1	0	16	0	1	18	1	0	5	4	12

SUM	303	296	12	7	277	11	26	301	48	2	149	130	154

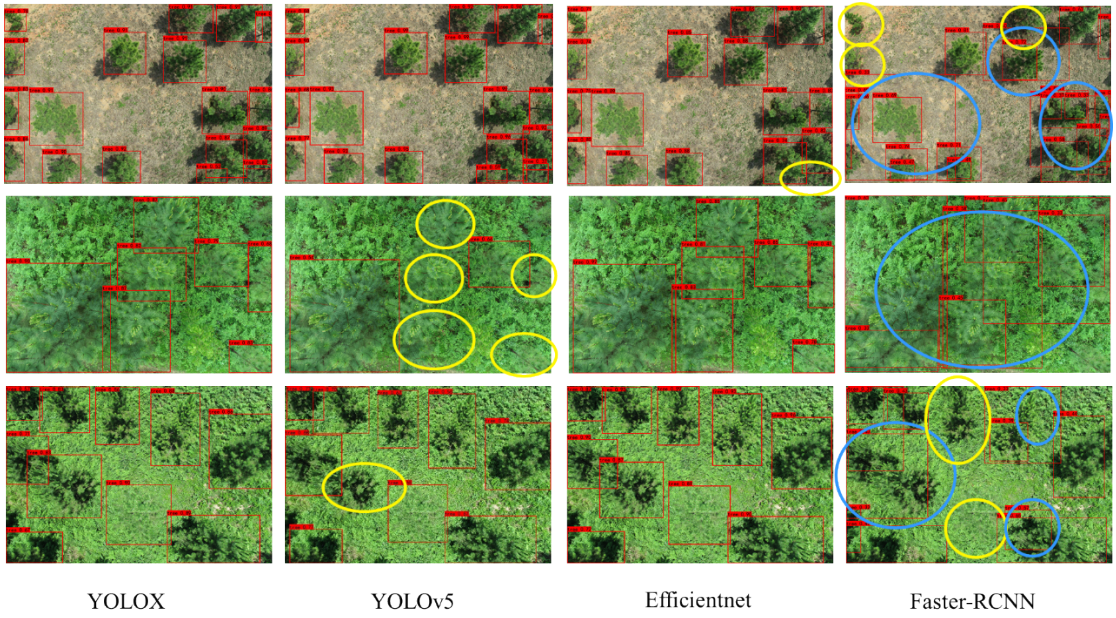


Figure 5 Detection effect of different models. The missed detections of the model are indicated by yellow circles and false detections are indicated by blue circles.

3.2 New shoots detection model

The balanced OT of the original loss and the unbalanced OT framework in this study were compared. We evaluated the different transposition cost matrices (including Euclidean distance L_{ij}, squared Euclidean distance $L^2{}_{ij}$ and exponential Euclidean distance $e^{L_{ij}}$), and the effectiveness of the perspective factor value in the transmission cost matrix on MAE (Figure 6). The base cost is Euclidean distance $L_{ij} = ||x_i - y_i||_2$. As clearly shown, the exponential function outperforms the traditional cost function based on

Euclidean distance, while the perspective-guided model gets the best performance. This proves the efficacy of non-equilibrium OT for density regression problems. This directly draws the point-by-point loss and pixel loss to penalize additional/missing density, while the original loss requires additional counting loss, reducing counting efficiency.

In addition, the TV loss in the original loss is supervised pixel by pixel by using the normalized dot plot, which is prone to overfitting. The loss we used was more efficient at pushing density from the background to the annotation than the square Euclidean cost. The blue section represents the results of the original loss function of CCTrans, while the orange section represents the results of our improved loss function in SPSC-net.

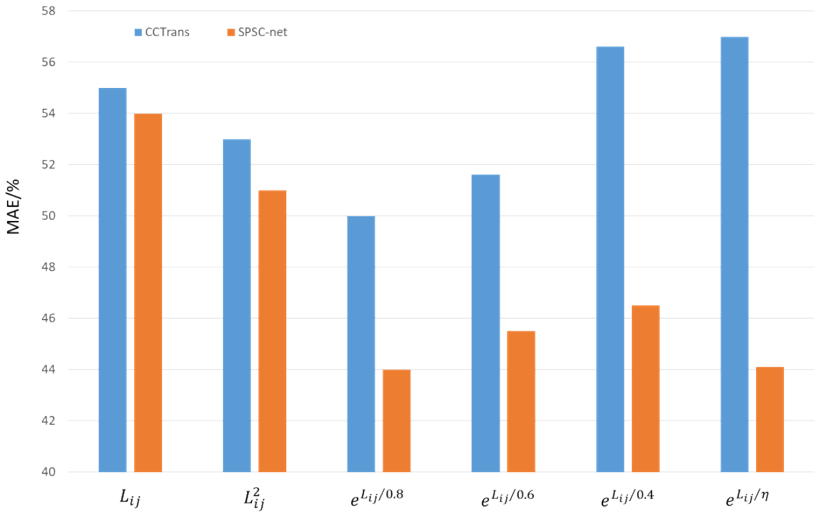


Figure 6 Comparison of different loss functions. The *x-axis* represents different cost functions, while the y-axis indicates the MAE results of the experiments.

The SPSC-net model had the lowest MSE and MAE among all the models, with values of 7.0 and 2.27- respectively (Table 3). Followed by the DM-Count and CSR-net, MCNN yielded the highest MSE and MAE results (34.50 and 30.76 respectively). Figure 7 illustrates the counting performance of SPSC-net, DM-Count, CSR-net and MCNN on the test set through a scatter plot. The results demonstrate that SPSC-Net achieves high prediction accuracy, followed by DM-Count, while MCNN has the poorest prediction performance. DM-Count and CSR-Net often overestimate the count, leading to predicted values higher than the true values.

Table 3 Comparison of new shoots counting of different methods

Model	MSE	MAE
SPSC-net	7.00	2.27
DM-Count	12.85	6.84
CSR-net	27.06	6.50
MCNN	31.12	30.76

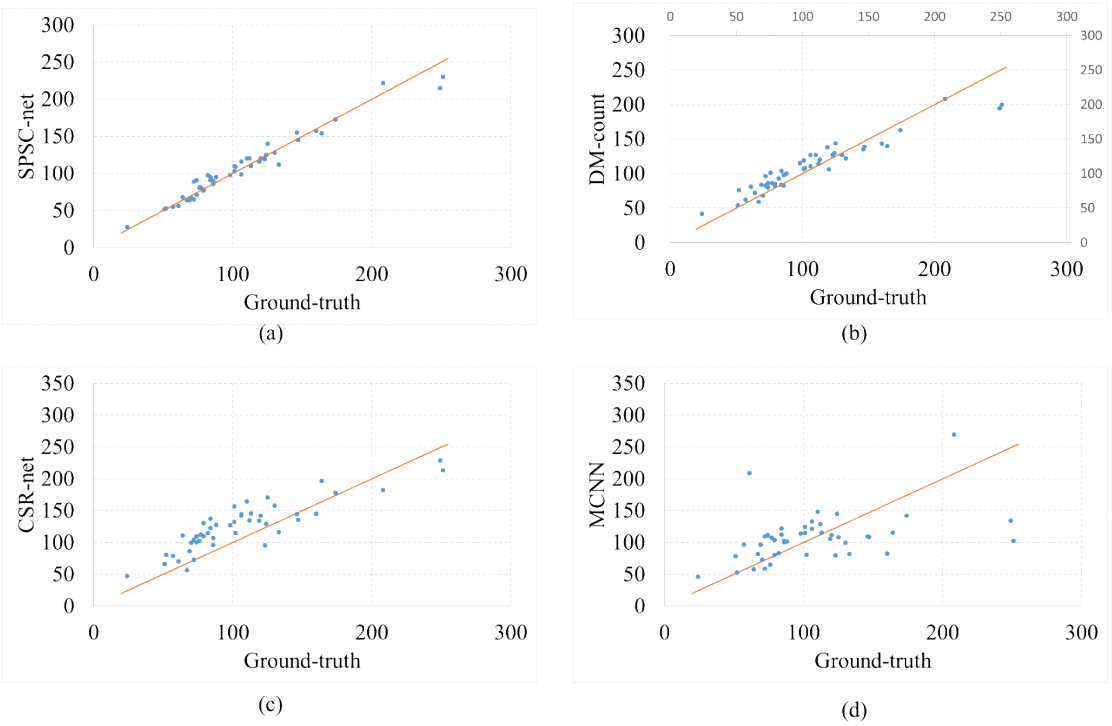


Figure 7 Counting results by different methods on test set. The x-axis showing the ground truth values and the y-axis indicating the predicted values.

The performance of SPSC-net, DM-Count, CSR-net and MCNN was visualized in Figure 8. The predicted density map of different models based on the same image and the predicted number of new shoots, compared with the original image, showed that the experimental image using the SPSC-net model can clearly reflect the distribution of the drawbacks and its denseness (Figure 8).

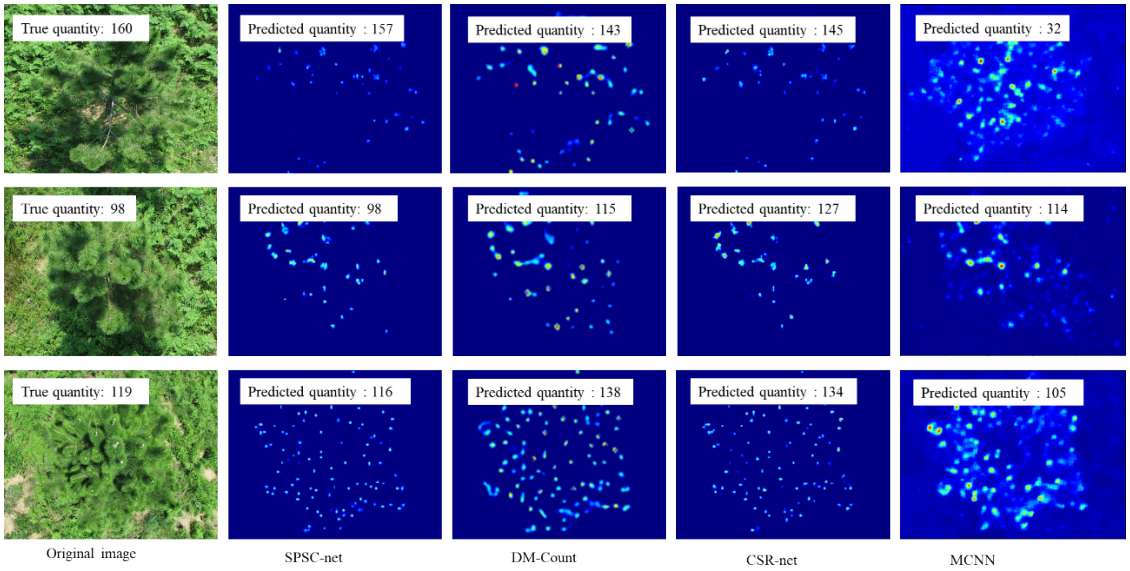


Figure 8 Density plot examples by different methods

3.3 Development of CountShoots application

In order to maximize the application value of the above model, we proposed to integrate and develop a slash pine extraction and new shoots counting system called CountShoots for forestry researchers. The system was implemented based on the Flask framework (Aggarwal 2014). By loading the trained object detection model and counting model, the individual crown extraction and automatic counting of slash pine are sequentially realized. The functional modules of the system has been separated into user interaction module, model loading module, plant extraction model and shoots counting module:

(1) User interaction module, which is used to submit the image to be predicted and display the counting result.

(2) The model loading module, loads the trained slash pine plant extraction and counting models into memory for the next step of plant extraction and new shoots counting.

(3) Plant extraction module, which extracts individual plants from the original UAV image using the trained plant extraction model (section 2.4).

(4) The shoots counting module, which uses the shoots counting model to count the new shoots of individual slash pine.

The system was operated in “request-response” mode. The process of new shoots counting is depicted in Figure 9. The specific process is as follows:

(1) Image select and upload. The dialog function is used to obtain the image path. The image is read and displayed on the front-end webpage through the “imshow” function. The user clicks the "Upload" button to send the image to the server.

(2) Plant extraction. This step is the first to be executed after the user triggers the 'counting' task. The system updates the trained YOLOX model (Section 2.4) to perform the canopy detection task and extracts the single

plant according to the detection results. The detection results are feedback to the front-end webpage.

(3) Shoots counting. This step is executed immediately after the second step. The system reads in the single plant image obtained in the second step, then performs the trained counting model SPSC-net (Section 2.5) to count the number of new shoots from the extracted canopy.

(4) Results feedback. The final outcome (i.e. the counting results) is feedbacked to the user interface.

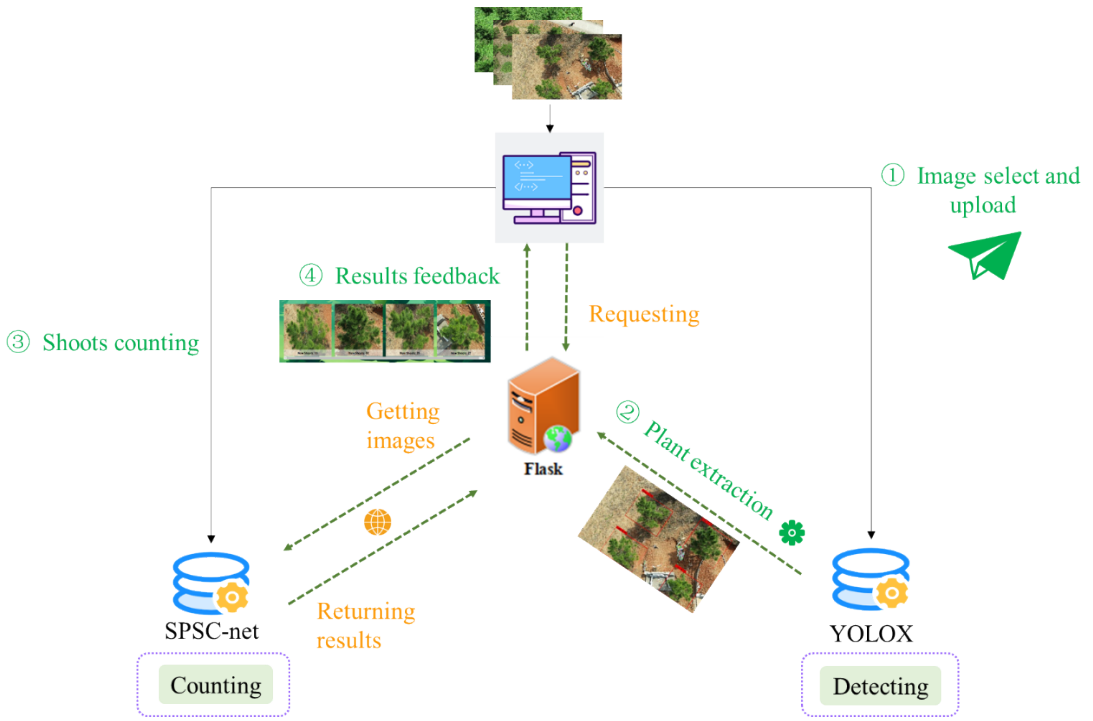
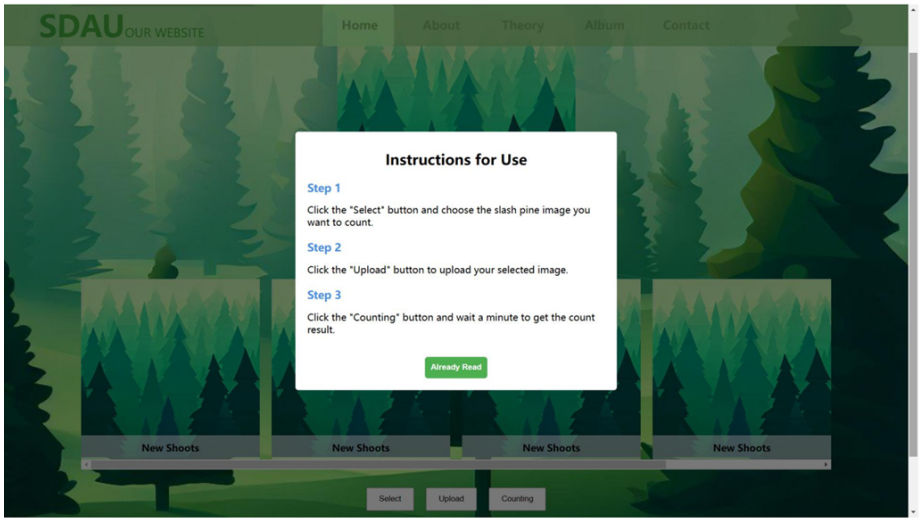


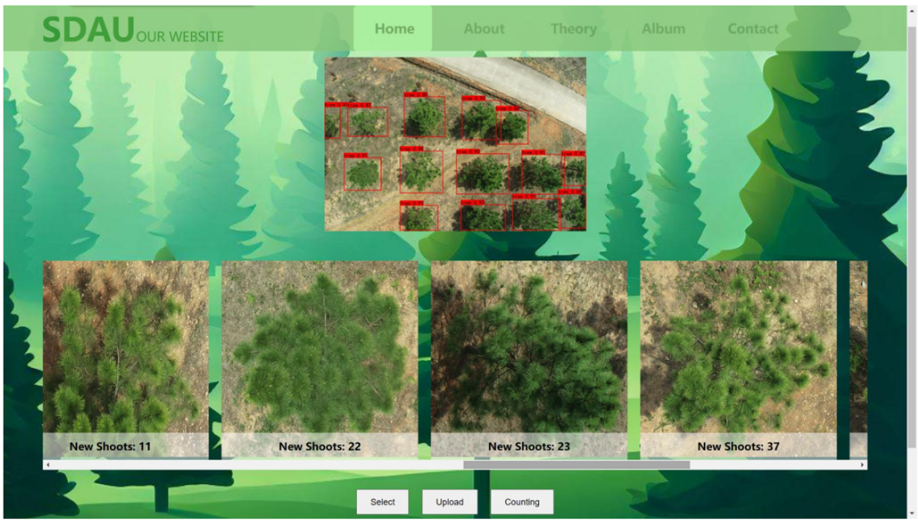
Figure 9 General flow chart of CountShoots

The counting interface of CountShoots is shown in Figure 10. After entering the counting page, new users can follow the instructions on the

homepage to perform counting tasks (Figure 10(a)). In the new shoots counting interface, users can click the “Select” button to specify image path, click "Upload" button to load the slash pine image and click the "Counting" button to count the number of new shoots. The users can directly view the number of new shoots in each crown, as shown in Figure 10(b).



(a)



(b)

Figure 10 The counting interface of CountShoots

4 Discussion

In this study, the slash pine trees in a breeding population were employed for the development of tree identification and new shoots detections model using UAV imaging. The experimental results demonstrate the superiority of the proposed framework on multi-scale image processing of slash pine. In addition, user-friendly software was created which makes it more convenient and suitable for the tree breeding or ecological researchers to be used.

Firstly we used a YOLOX model to identify the tree species and location from the complex background RGB images and compared it with other state-of-art published models. The results show that our model has the highest detection accuracy compared to YOLOv5, Efficientnet and Faster-RCNN models. Similar results have been found in other object detection studies (Yang, et al. 2022,Zhang, et al. 2022). YOLOv5 showed less accuracy on the object detection than YOLOX model (Ge, et al. (2022), with values of 86.42% and 83.95% on object detection in the Beijing capital airport in China using the Synthetic Aperture Radar (SAR) imagery, but less than the spatial orientation attention enhancement network

(SOAEN) (90.12%). However, this result is lower than our study (i.e. recognition accuracy of 95.47% with a threshold equal to 0.75). Our YOLOX model produced a similar AP to the results of Wu, et al. (2022). YOLOX and Tree Crown Detection network (TCDnet) yielded accuracy rates of 96.30% and 96.76% respectively based on the identification of bayberry trees using UAV optical imagery. YOLOX model shows the best performance in the case of target coincidence, blurry and complex background. It innovatively employed the decoupled head and introduces the anchor-free detector to improve the performances of detection for the YOLOX model which can accurately find the overlap and small target of slash pine trees.

In the new shoot counting model, the newly developed counting network SPSC-net performed better than DM-Count, CSR-net and MCNN models. SPSC-net uses the self-attention mechanism to capture the global features of the new shoots and designs the feature pyramid fusion module to supplement the details with the underlying features to obtain the extractive features rich in semantics, detail information and global features. The hollow convolutional network was used in SPSC-net to better capture

image information at different scales and predict the number of new shoots more accurately. In addition, in the SPSC-net, the unbalanced OT loss is designed to replace the OT loss in the CCTrans model. Firstly, the method directly takes the use of point-by-point loss and pixel loss to penalize the additional/missing density, while the original loss requires additional counting loss, which reduces counting efficiency. Secondly, the TV loss in the original loss is supervised pixel by pixel using the normalized dot plot, which is prone to overfitting. Through our experiment, we proved that the loss we used was more efficient at pushing density from the background to the annotation than the square Euclidean cost. The SPSC-net counting model performs better than the YOLOv5-SBiC model built by Liang, et al. (2023) which yields a recognition accuracy value of 79.6% on the detection of late-autumn shoots of Lichi trees, and the model built by Hong, et al. (2022) on the detection of wheat ear fusarium head blight based on RGB images, with an accuracy of 93.69%. However, it is important to note that this comparison serves as a rough overview and does not possess the same level of rigor as the main results.

It should also be mentioned that limitations occur in our model. Firstly,

slash pine trees are a type of macrophanerophytes with tall crowns and luxuriant branches, and the new shoots at the lower layer may be shielded by the upper layers of the canopy, which cannot be captured by the UAV and detected by the models, making it difficult to accurately identify all the new shoots on the slash pine trees. The number of the lower layers of new shoots overlapped is much less and most of these are small since they are at the early growth stage. Therefore, these have little effect on the total among new shoots. Secondly, the flying height of the UAV adopted to capture remote sensing images is strictly limited to 10-15 m. A higher flight height potentially reduces the detection accuracy due to the low resolution of remote sensing images. The impact of flight altitude and image resolution on the accuracy of new shoot counting was not investigated in this study, which could be another research topic for future studies. To acquire UAV imagery with limited flight height, it is time-consuming to survey for all trees in large orchards. Further testing is required for the practical applications of the model in other pine plantations.

5 Conclusion

To the best of our knowledge, our study developed a pipeline identifying tree crowns and counting new shoots using two separate models: the SPSC-net slash pine new shoots counting model and the tree crown detection network YOLOX. We obtained an accurate counting of new shoots of a single slash pine tree. The extraction network was implemented for the derivation and labeling of slash pine, reducing the influence of complex backgrounds such as soil and light. The counting network deployed used multi-void rate convolution to fuse multi-scale features, increasing the accuracy of tip positioning by using non-equilibrium transmission and perspective-guided loss, generating a high-resolution density map of new shoots distribution, which provides a better supporting tool for forestry researchers. Meanwhile, an automatic new shoot counting system for slash pine called CountShoots was built, which efficiently and accurately estimated the number and distribution density of slash pine new shoots and provided reliable data support for subsequent genetic breeding and efficient breeding research of slash pine.

Data Availability

The data in this study are available on upon request from the corresp onding author. The counting system were publicly released at https://g ithub.com/haohuihui5019/CountShoots.

Authors' contributions

Xia Hao conducted the experiments and wrote the manuscript. **Yue Cao** provided the data collection, labeling and analysis. **Zhaoxu Zhang** constructed and tested the application system. **Yanjie Li** designed the study, supported the data collection and field experiments and revised the manuscript. **Federico Tomasetto, Weiqi Yan and Cong Xu** performed the revisions of the manuscript, and all authors read and approved the final manuscript.

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 study.

Acknowledgments

This research was supported by Fundamental Research Funds of CAF, No. CAFYBB2022QA001 and the Zhejiang Science and Technology Major

Program on Agricultural New Variety Breeding, No. 2021C02070-7-3.

References:

1. Ding, X., Diao, S., Luan, Q., Wu, H.X., Zhang, Y. and Jiang, J. 2022 A transcriptome-based association study of growth, wood quality, and oleoresin traits in a slash pine breeding population. *PLoS genetics*, **18** (2), e1010017.

2. Lai, M., Zhang, L., Lei, L., Liu, S., Jia, T. and Yi, M. 2020 Inheritance of resin yield and main resin components in Pinus elliottii Engelm. at three locations in southern China. *Industrial crops and products*, **144**, 112065.

3. FIFE, D.N. and Nambiar, E. 1982 Accumulation and retranslocation of mineral nutrients in developing needles in relation to seasonal growth of young radiata pine trees. *Annals of Botany*, **50** (6), 817-829.

4. Kellomäki, S. and Strandman, H. 1995 A model for the structural growth of young Scots pine crowns based on light interception by shoots. *Ecological Modelling*, **80** (2-3), 237-250.

5. Fife, D. and Nambiar, E. 1984 Movement of nutrients in radiata pine needles in relation to the growth of shoots. *Annals of Botany*, **54** (3), 303-314.

6. Whitmore, F. and Zahner, R. 1966 Development of the xylem ring in stems of young red pine trees. *Forest Science*, **12** (2), 198-210.

7. Li, Z., Guo, R., Li, M., Chen, Y. and Li, G. 2020 A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, **176**, 105672.

8. Zweifel, R., Etzold, S., Sterck, F., Gessler, A., Anfodillo, T., Mencuccini, M. *et al.* 2020 Determinants of legacy effects in pine trees–implications from an irrigation-stop experiment. *New Phytologist*, **227** (4), 1081-1096.

9. Stadler, B., Müller, T., Orwig, D. and Cobb, R. 2005 Hemlock woolly adelgid in New England forests: canopy impacts transforming ecosystem processes and landscapes. *Ecosystems*, **8**, 233-247.

10. Rosati, A., Paoletti, A., Al Hariri, R. and Famiani, F. 2018 Fruit production and branching density affect shoot and whole-tree wood to leaf biomass ratio in olive. *Tree Physiology*, **38** (9), 1278-1285.

11. Dalla Corte, A.P., Rex, F.E., Almeida, D.R.A.d., Sanquetta, C.R., Silva, C.A., Moura, M.M. *et al.* 2020 Measuring individual tree diameter and height using GatorEye High-Density UAV-Lidar in an integrated crop-livestock-forest system. *Remote Sensing*, **12** (5), 863.

12. Picos, J., Bastos, G., Míguez, D., Alonso, L. and Armesto, J. 2020 Individual tree detection in a eucalyptus plantation using unmanned aerial vehicle (UAV)-

LiDAR. *Remote Sensing*, **12** (5), 885.

13. López-Granados, F., Torres-Sánchez, J., Jiménez-Brenes, F.M., Arquero, O., Lovera, M. and de Castro, A.I. 2019 An efficient RGB-UAV-based platform for field almond tree phenotyping: 3-D architecture and flowering traits. *Plant Methods*, **15** (1), 1-16.

14. Xie, C. and Yang, C. 2020 A review on plant high-throughput phenotyping traits using UAV-based sensors. *Computers and Electronics in Agriculture*, **178**, 105731.

15. Pound, M.P., Atkinson, J.A., Wells, D.M., Pridmore, T.P. and French, A.P. Deep learning for multi-task plant phenotyping, pp. 2055-2063.

16. Sun, J., Yang, K., Chen, C., Shen, J., Yang, Y., Wu, X. *et al.* 2022 Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network. *Computers and Electronics in Agriculture*, **193**, 106705.

17. Bayraktar, E., Basarkan, M.E. and Celebi, N. 2020 A low-cost UAV framework towards ornamental plant detection and counting in the wild. *ISPRS Journal of Photogrammetry and Remote Sensing*, **167**, 1-11.

18. Su, J., Liu, C., Coombes, M., Hu, X., Wang, C., Xu, X. *et al.* 2018 Wheat yellow

rust monitoring by learning from multispectral UAV aerial imagery. *Computers and electronics in agriculture*, **155**, 157-166.

19. Barrero, O. and Perdomo, S.A. 2018 RGB and multispectral UAV image fusion for Gramineae weed detection in rice fields. *Precision Agriculture*, **19** (5), 809-822.

20. Evgeniou, T. and Pontil, M. Support vector machines: Theory and applications. Springer, pp. 249-257.

21. Breiman, L. 2001 Random Forests. *Machine Learning*, **45** (1), 5-32.

22. Singh, A.K., Ganapathysubramanian, B., Sarkar, S. and Singh, A. 2018 Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in plant science*, **23** (10), 883-898.

23. van Dijk, A.D.J., Kootstra, G., Kruijer, W. and de Ridder, D. 2021 Machine learning in plant science and plant breeding. *Iscience*, **24** (1), 101890.

24. Chlingaryan, A., Sukkarieh, S. and Whelan, B. 2018 Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, **151**, 61-69.

25. Li, H., Wang, P. and Huang, C. 2022 Comparison of Deep Learning Methods for Detecting and Counting Sorghum Heads in UAV Imagery. *Remote Sensing*,

14 (13), 3143.

26. LeCun, Y., Bengio, Y. and Hinton, G. 2015 Deep learning. *Nature*, **521** (7553), 436-444.

27. Bengio, Y., Courville, A. and Vincent, P. 2013 Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35** (8), 1798-1828.

28. Kamilaris, A. and Prenafeta-Boldú, F.X. 2018 Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, **147**, 70-90.

29. Chen, Y., Li, W., & Sakaridis, C. 2019 Multi-scale fusion with CNNs for enhancing crowd counting. *Pattern Recognition*, **90**, 119-130.

30. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. *et al.* 2020 An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

31. Liang, D., Chen, X., Xu, W., Zhou, Y. and Bai, X. 2022 TransCrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, **65** (6), 1-14.

32. Sun, G., Liu, Y., Probst, T., Paudel, D.P., Popovic, N. and Van Gool, L. 2021 Boosting crowd counting with transformers. *arXiv preprint arXiv:2105.10926*.

33. Tian, Y., Chu, X. and Wang, H. 2021 Cctrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483*.

34. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X. *et al.* 2021 Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, **34**, 9355-9366.

35. Song, Z., Tomasetto, F., Niu, X., Yan, W.Q., Jiang, J. and Li, Y. 2022 Enabling Breeding Selection for Biomass in Slash Pine Using UAV-Based Imaging. *Plant Phenomics*, **2022**.

36. Tzutalin, D. 2015 LabelImg. *GitHub repository*, **6**.

37. Yin, X., Chen, S., Hu, E. and Zhang, D. 2010 Semi-supervised clustering with metric learning: An adaptive kernel method. *Pattern Recognition*, **43** (4), 1320-1333.

38. Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J. 2021 Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

39. Redmon, J. and Farhadi, A. 2018 Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

40. Zhu, M., Han, K., Yu, C. and Wang, Y. 2020 Dynamic feature pyramid networks for object detection. *arXiv preprint arXiv:2012.00779*.

41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. *et al.* Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008.

42. Wang, B., Liu, H., Samaras, D. and Nguyen, M.H. 2020 Distribution matching for crowd counting. *Advances in neural information processing systems*, **33**, 1595-1607.

43. Özleyen, S.Y. and Aptoula, E. Crowd Counting with Distribution Matching and Dilated Networks. IEEE, pp. 1-4.

44. Li, Y., Zhang, X. and Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, pp. 1091-1100.

45. Yang, L., Yuan, G., Zhou, H., Liu, H., Chen, J. and Wu, H. 2022 RS-YOLOX: A High-Precision Detector for Object Detection in Satellite Remote Sensing Images. *Applied Sciences*, **12** (17), 8707.

46. Zhang, Y., Zhang, W., Yu, J., He, L., Chen, J. and He, Y. 2022 Complete and accurate holly fruits counting using YOLOX object detection. *Computers and Electronics in Agriculture*, **198**, 107062.

47. Ge, J., Wang, C., Zhang, B., Xu, C. and Wen, X. 2022 Azimuth-Sensitive Object Detection of High-Resolution SAR Images in Complex Scenes by Using

a Spatial Orientation Attention Enhancement Network. *Remote Sensing*, **14** (9), 2198.

48. Wu, W., Fan, X., Qu, H., Yang, X. and Tjahjadi, T. 2022 TCDNet: Tree Crown Detection From UAV Optical Images Using Uncertainty-Aware One-Stage Network. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1-5.

49. Liang, J., Chen, X., Liang, C., Long, T., Tang, X., Shi, Z. *et al.* 2023 A detection approach for late-autumn shoots of litchi based on unmanned aerial vehicle (UAV) remote sensing. *Computers and Electronics in Agriculture*, **204**, 107535.

50. Hong, Q., Jiang, L., Zhang, Z., Ji, S., Gu, C., Mao, W. *et al.* 2022 A Lightweight Model for Wheat Ear Fusarium Head Blight Detection Based on RGB Images. *Remote Sensing*, **14** (14), 3481.

Inputs

Step 1:Single plant extraction

Data labeling

Object detection

Object extraction

Step2: Shoots counting

Outputs

Countings:33

Data labeling

Density map generation

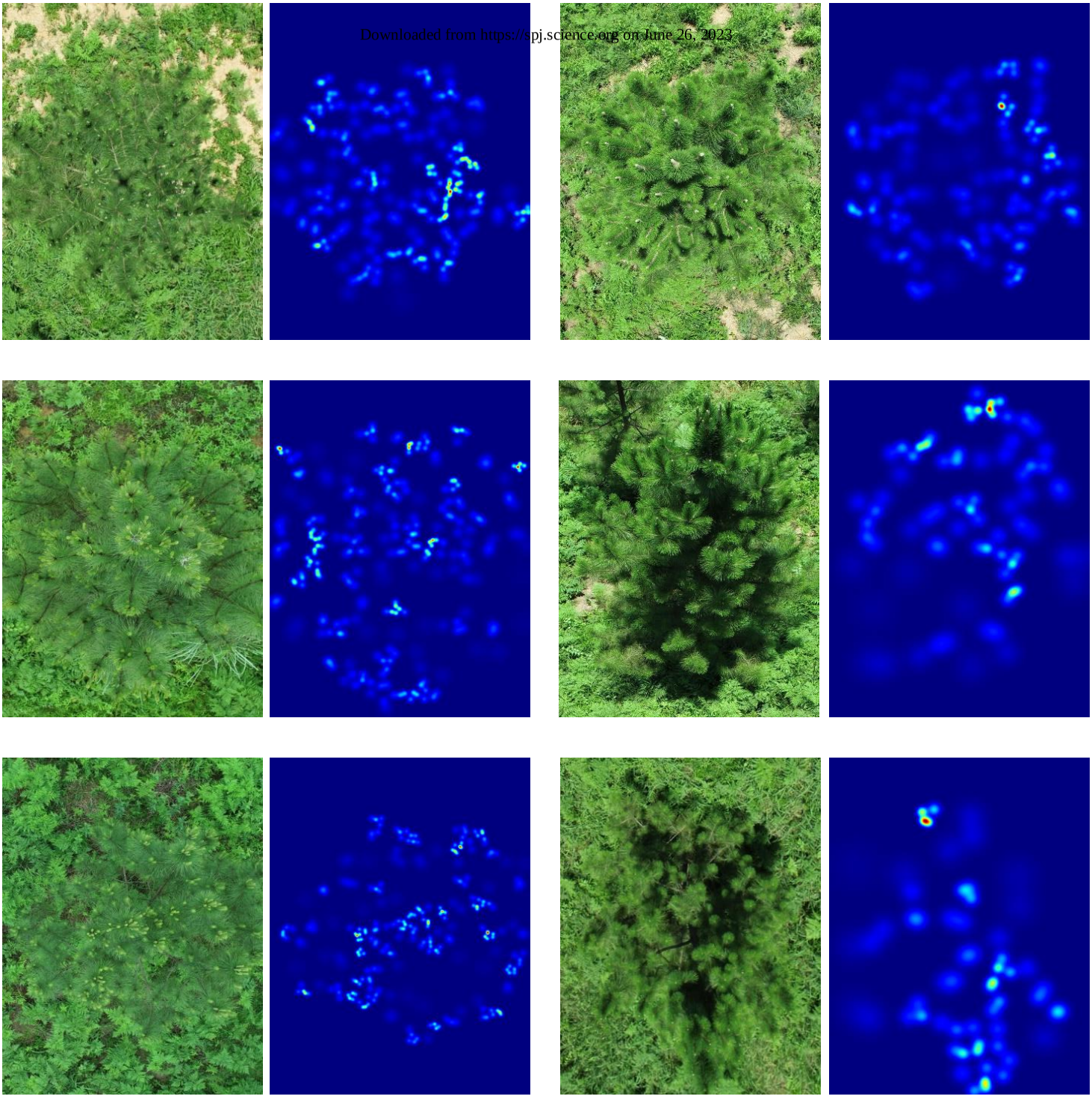
Shoots counting

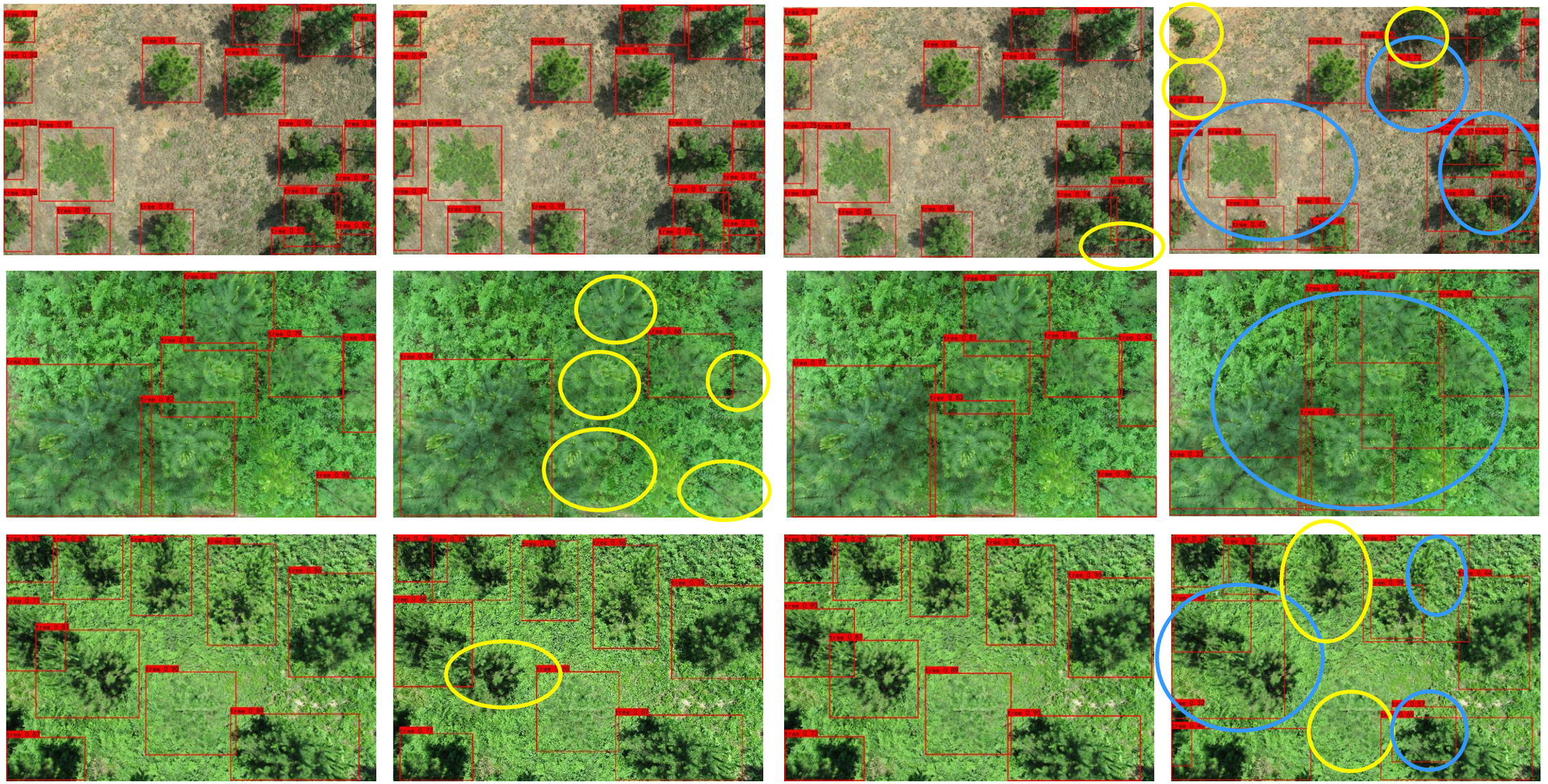
(a)

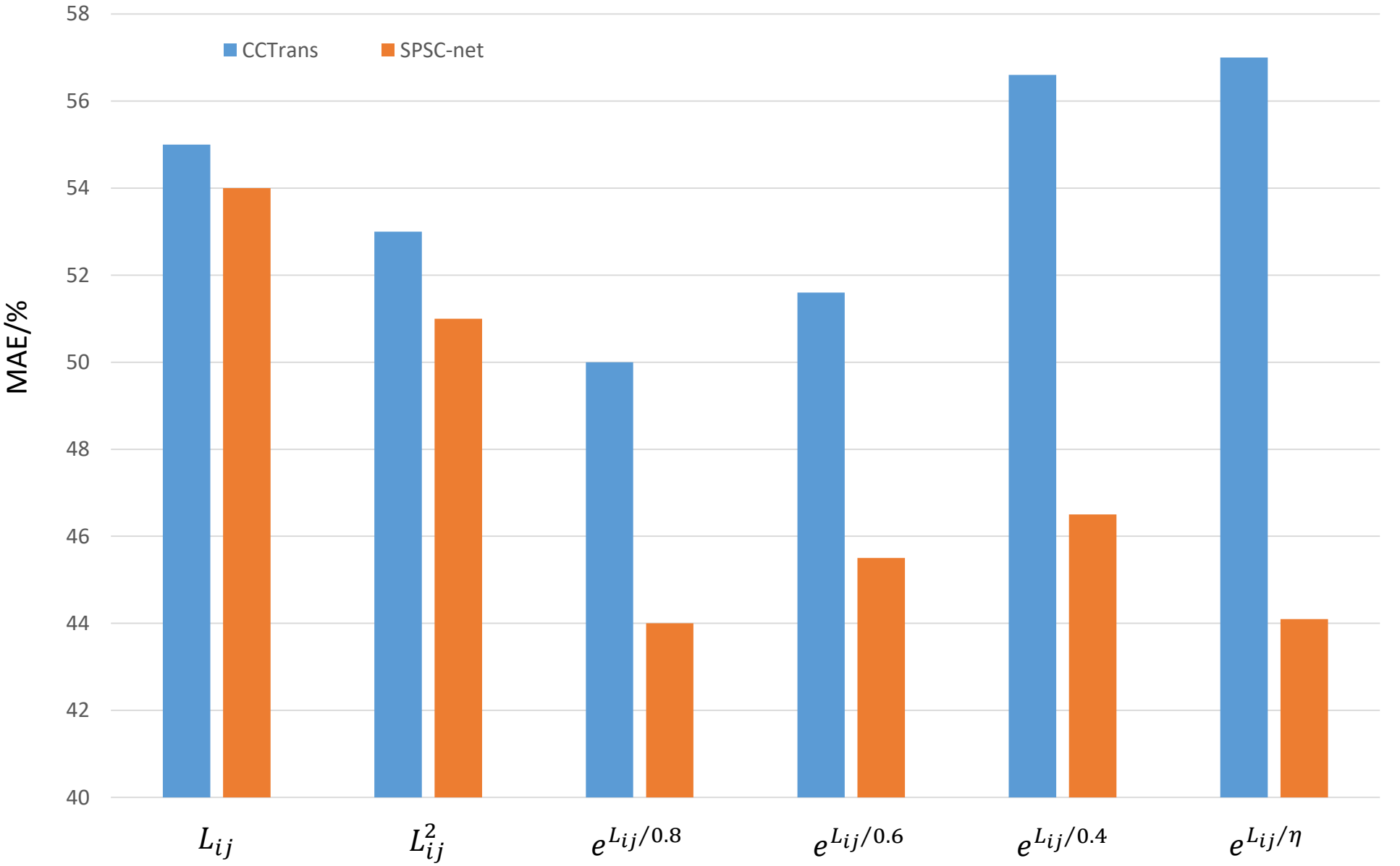
(b)

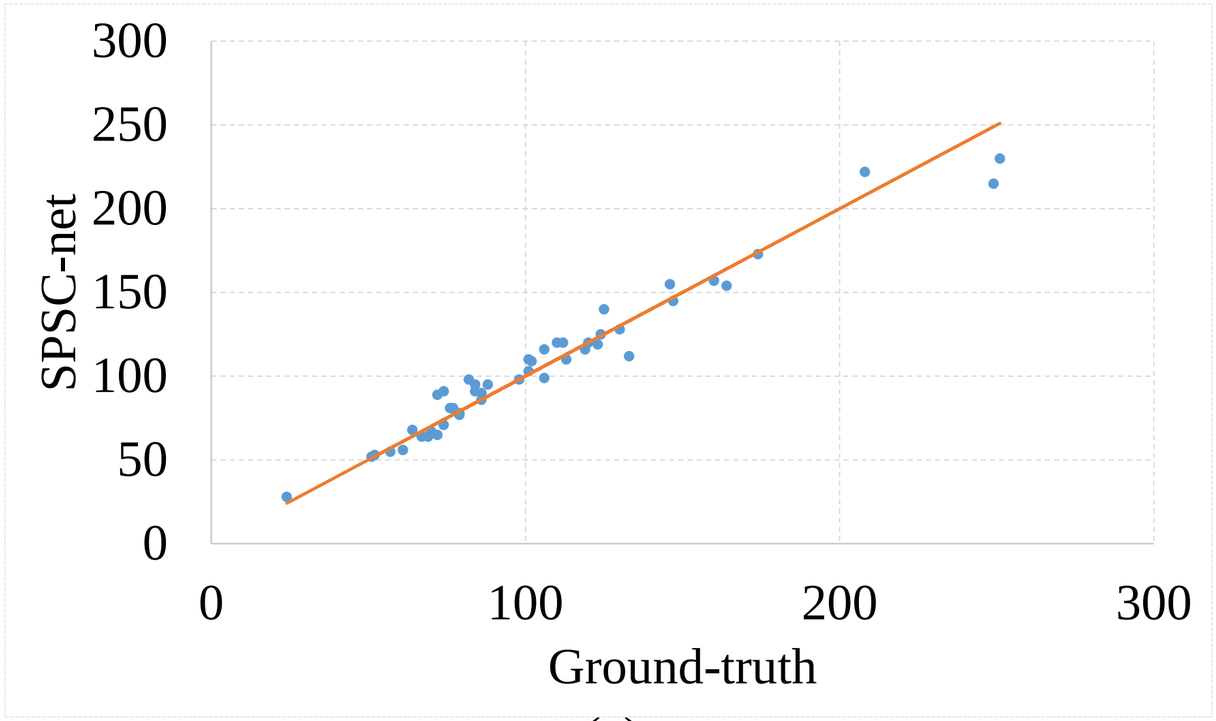
(a)

(b)

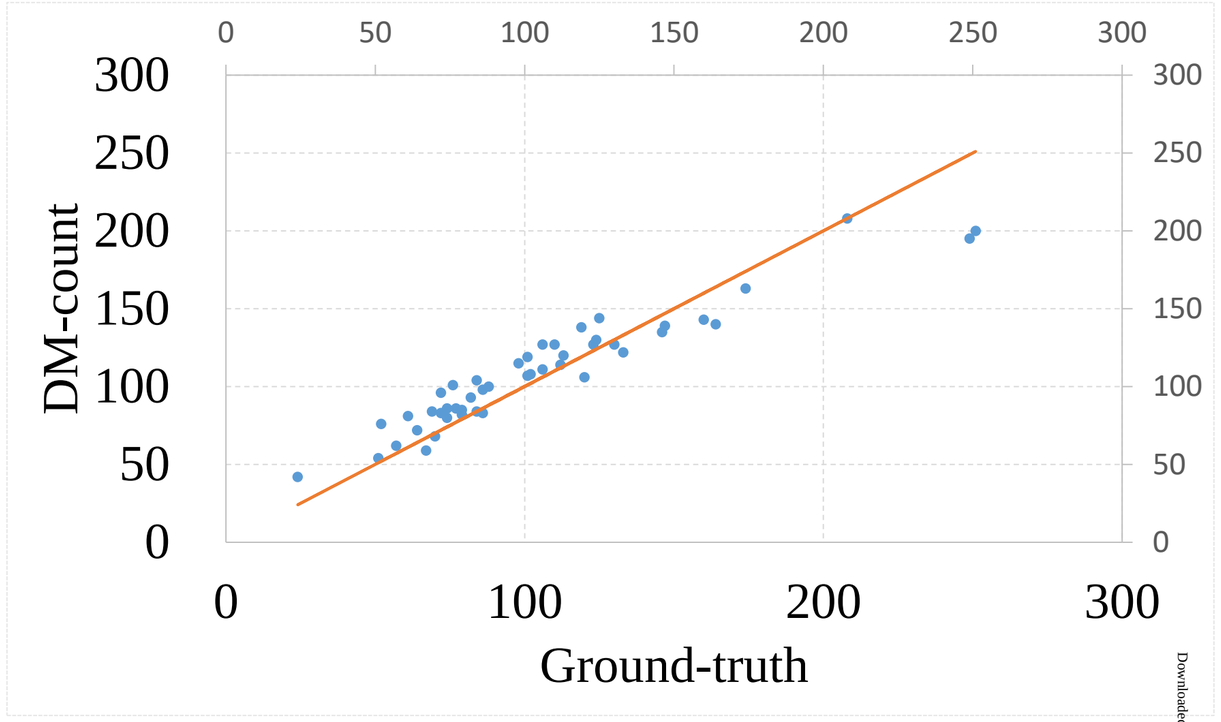




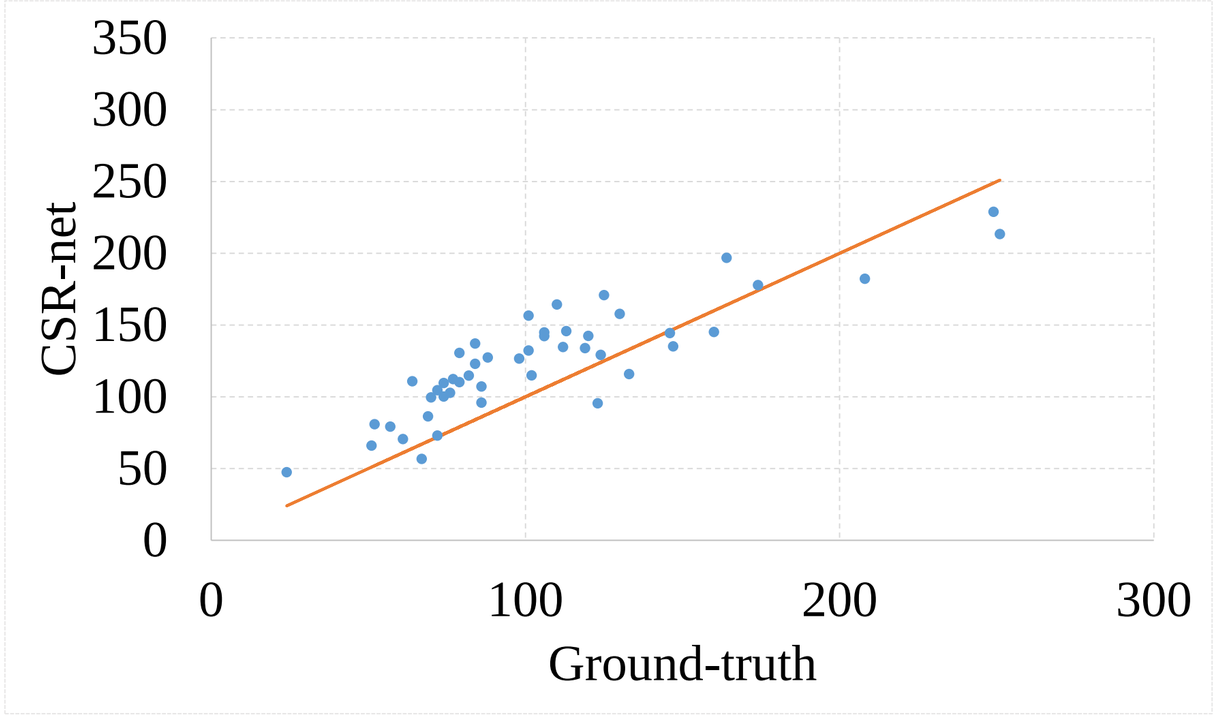




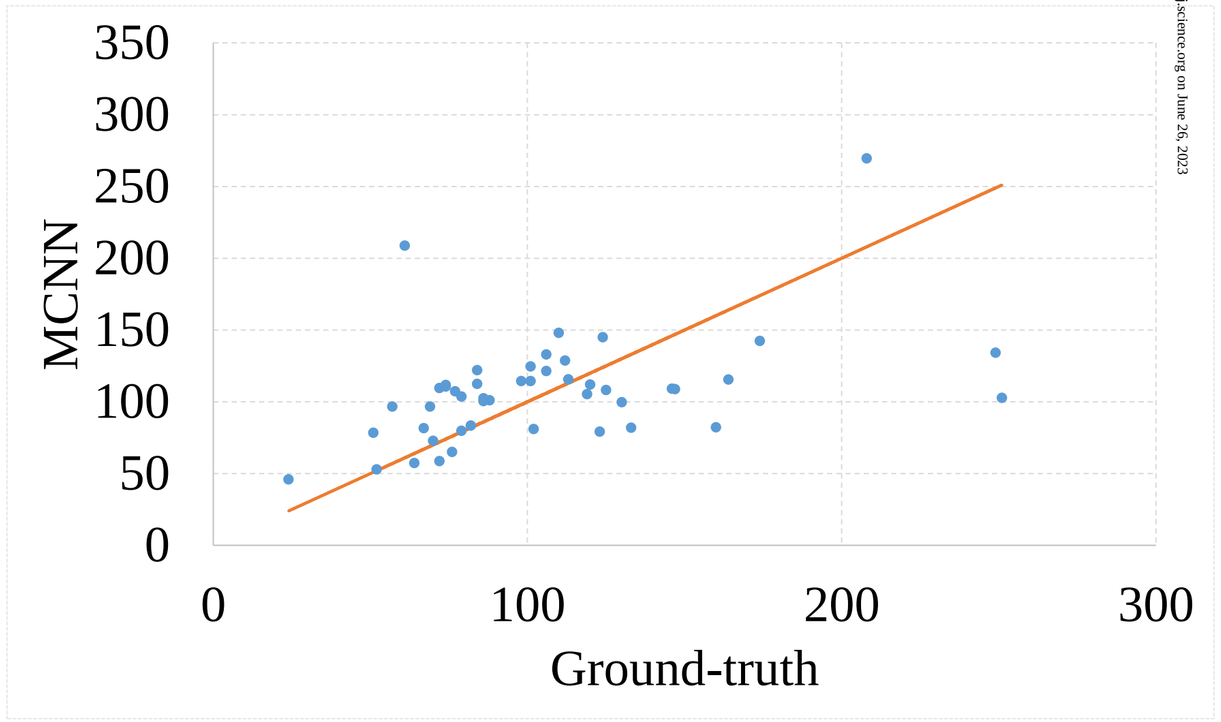
(a)



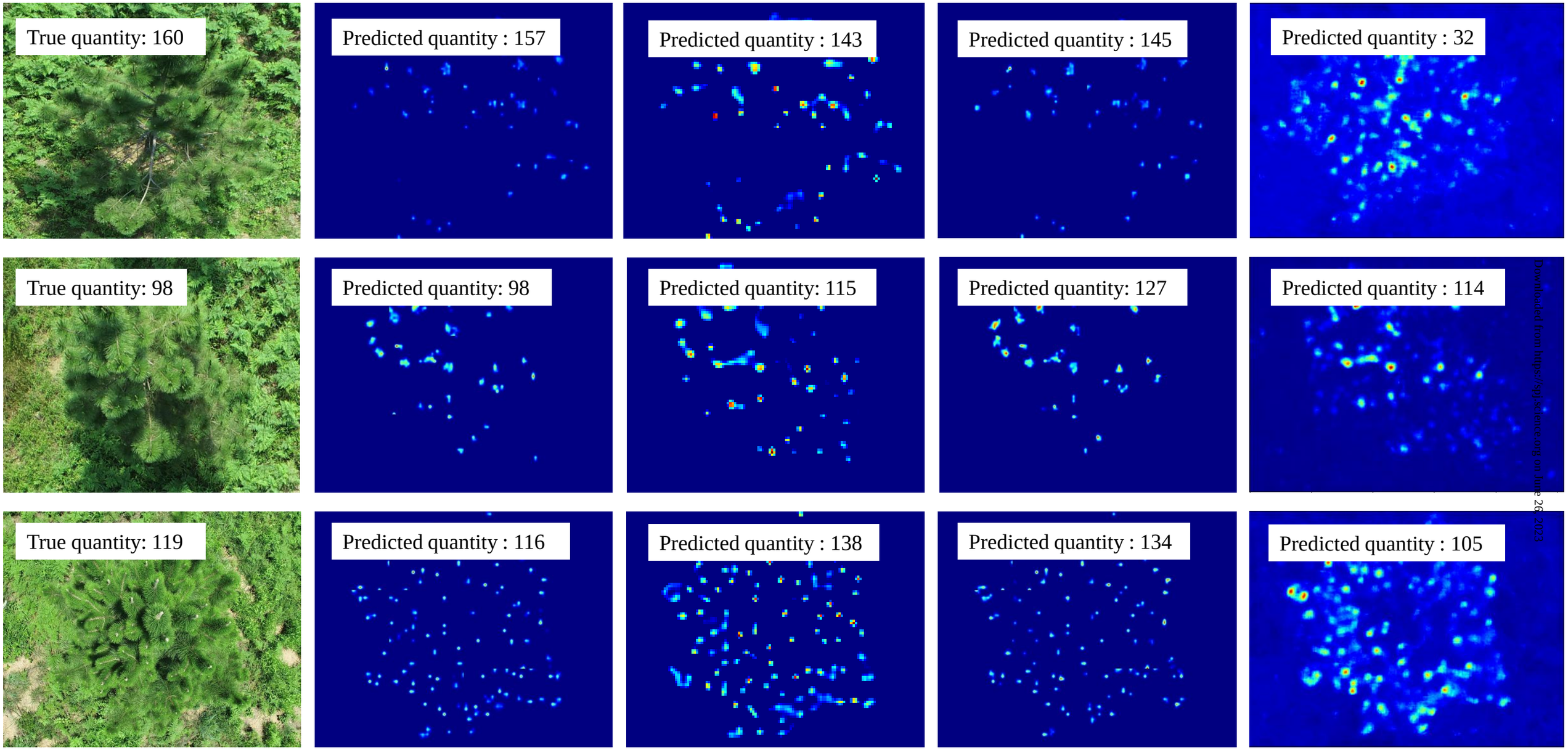
(b)

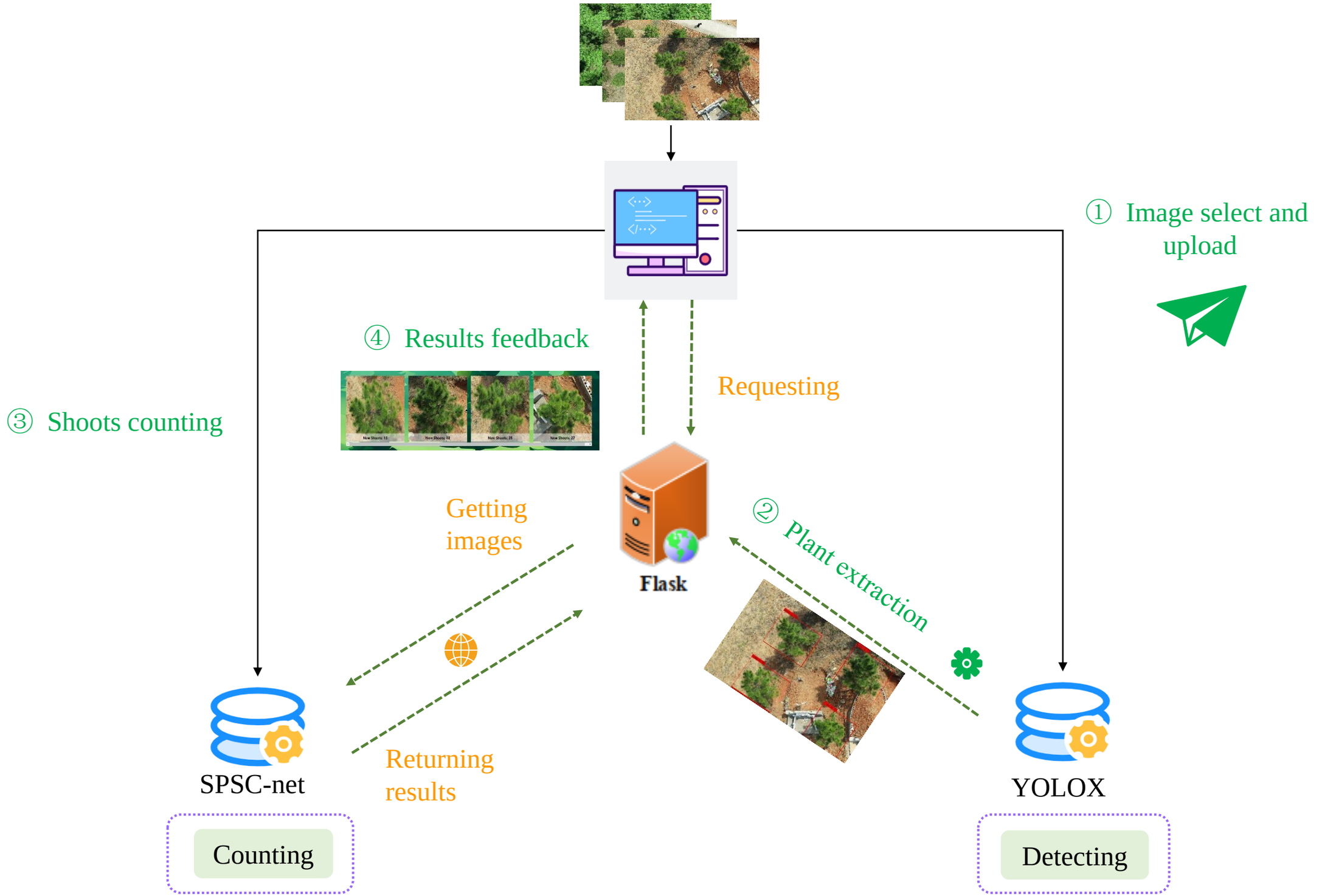


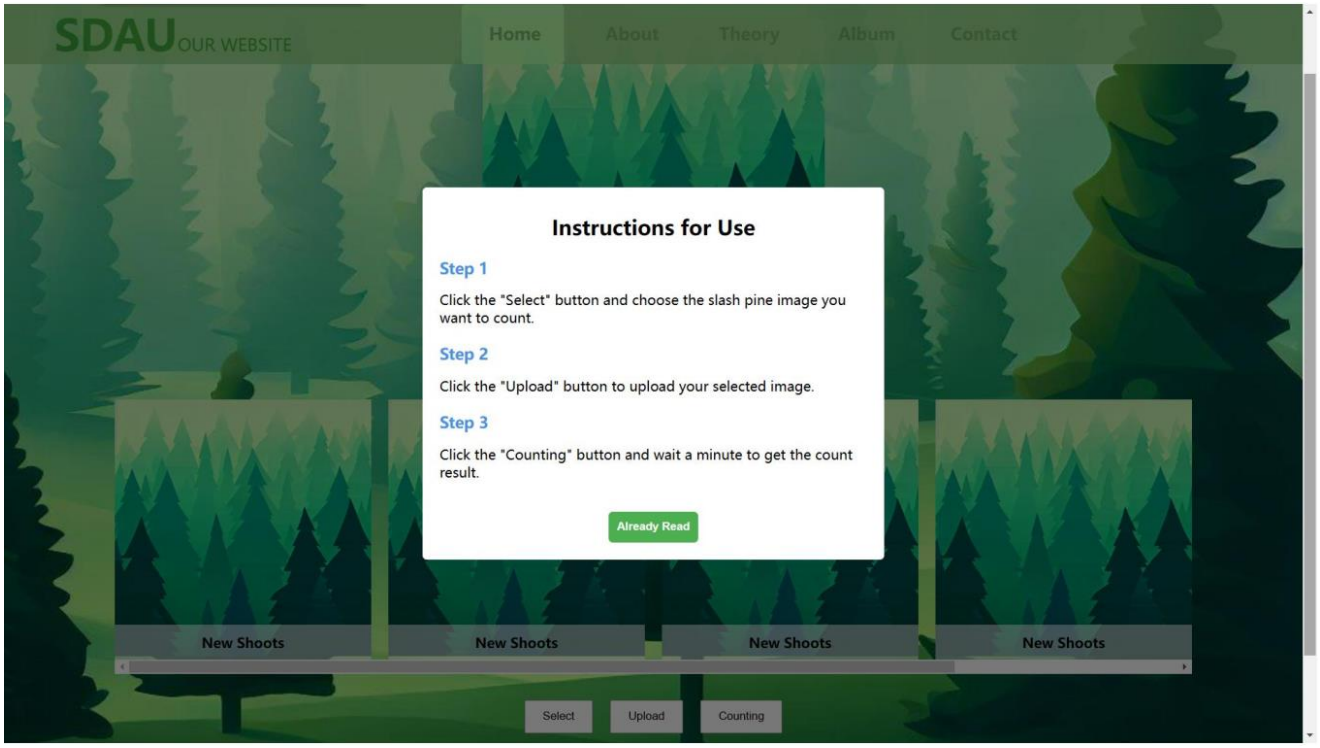
(c)



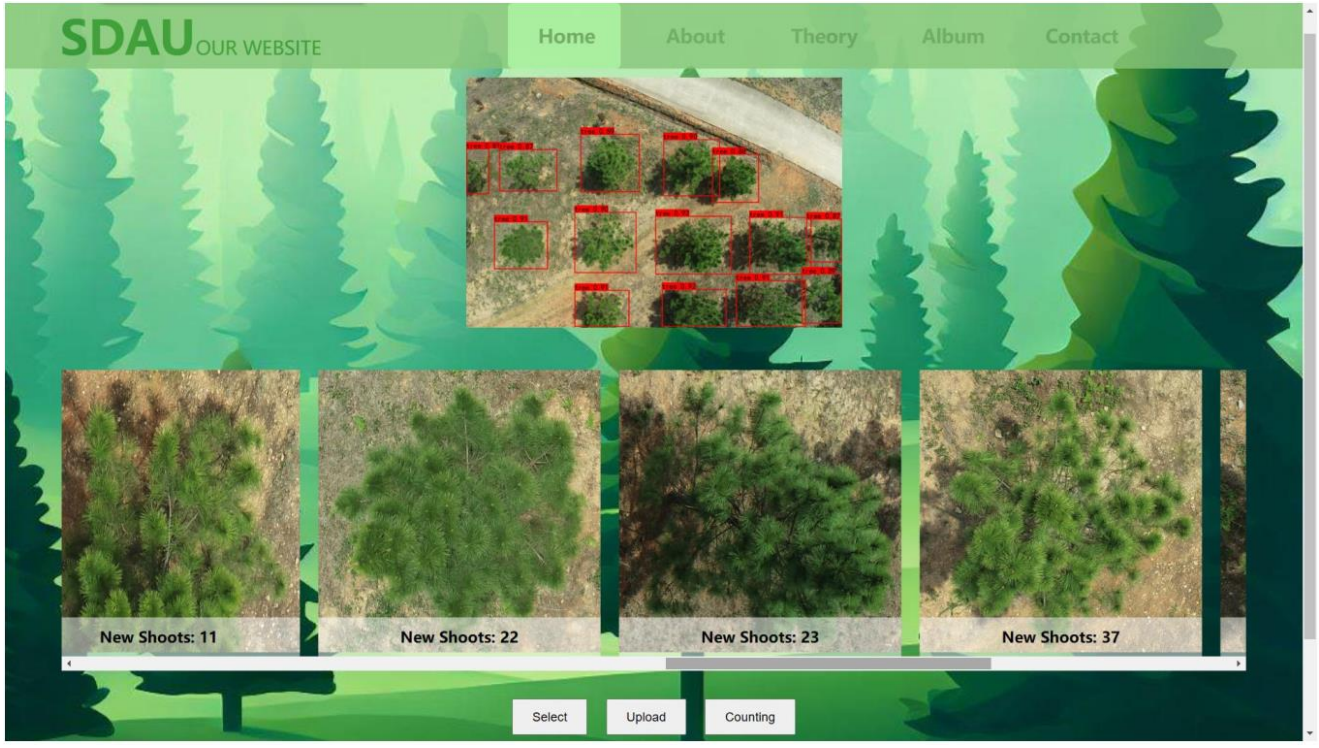
(d)







(a)



(b)