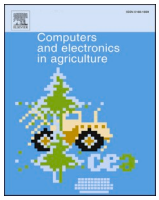
Computers and Electronics in Agriculture 209 (2023) 107822

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



A comparison between Pixel-based deep learning and Object-based image analysis (OBIA) for individual detection of cabbage plants based on UAV Visible-light images

Zhangxi Ye [a], Kaile Yang [a], Yuwei Lin [a], Shijie Guo [a], Yiming Sun [a], Xunlong Chen [a], Riwen Lai [a], Houxi Zhang [a,b,c,d,*]

[a] College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350028, China
[b] Key Laboratory of State Forestry and Grassland Administration on Soil and Water Conservation of Red Soil Region in Southern China, Fuzhou 350002, China
[c] Cross-Strait Collaborative Innovation Center of Soil and Water Conservation, Fuzhou 350002, China
[d] National Positioning Observation and Research Station of Red Soil Hill Ecosystem in Changting, Fuzhou 350002, China

ARTICLE INFO

Keywords:
Target detection
Plant count statistics
Deep learning
OBIA
Watershed algorithm

ABSTRACT

It is challenging to accurately and rapidly extract crops based on the ultra-high spatial resolution images of uncrewed aerial vehicle (UAV). Object-based image analysis (OBIA) was regarded as an effective technique for high-spatial-resolution image classification because of its ability to achieve high accuracy by integrating multi-dimensional features. In recent years, deep learning (DL) techniques, with their ability to automatically learn image features from a large number of images, have shown great potential for crop monitoring. However, a systematic comparison of these two mainstream methods for monitoring the crop phenotype has not been conducted. Therefore, this study compares the performance of two advanced methods, DL and OBIA, in individual cabbage plant detection tasks. The results show that the Mask R-CNN deep learning model outperforms the object-based image analysis-multilevel distance transform watershed segmentation (OBIA-MDTWS) method in crop extraction and counting, with an overall mean F1-Score, accuracy of 2.70, 4.15 percentage points higher, respectively. Moreover, the Mask R-CNN deep learning model has higher computing efficiency, which is 3.74 times higher than the OBIA-MDTWS model. In summary, this study shows that the Mask R-CNN deep learning model performs better in vegetable extraction and quantity estimation, providing technical support for subsequent field nursery management and fine planting.

1. Introduction

The demand for vegetables, being essential crops, has recently increased with the rapid growth of the global population (Jiang et al., 2019; Mavridou et al., 2019; Wang et al., 2021). Accurate and efficient vegetable extraction and detection are critical for managing the supply of water and fertilizers, optimizing vegetable production, and improving yield estimation accuracy in agriculture (Bian et al., 2022; Feng et al., 2020). However, because vegetables are easily affected by cultivation methods and growing environments, seedling emergence and growth stages greatly vary, which brings particular challenges to the identification and monitoring of vegetables (Ahansal et al., 2022). Therefore, the accurate phenotypic identification of crops in the field has become a research hotspot.

Current crop identification and monitoring methods mainly include ground surveys and remote-sensing-based methods (Fan et al., 2021; Kumari et al., 2022). Owing to the influence of human and material conditions, traditional manual methods have several shortcomings such as low efficiency, high cost, intense subjectivity, and low timeliness. Therefore, it is not suitable for crop extraction in large areas. Moreover, it is challenging to apply it to the periodic and continuous monitoring of crops (Jiang et al., 2022). Remote sensing has been indispensable in modern agriculture monitoring because it can obtain large-scale spatial farmland information and provide the necessary decision support for crop management. Satellite-based images are commonly used for macroscopic observations, and there have been many successful cases of the accurate identification and detection of extensive crops. However, satellite remote sensing has the disadvantages of low resolution, significant weather influence, and poor timeliness, making it difficult to obtain timely and effective data for small crop-monitoring tasks.

* Corresponding author.

https://doi.org/10.1016/j.compag.2023.107822
Received 25 August 2022; Received in revised form 28 March 2023; Accepted 30 March 2023
Available online 12 April 2023

Uncrewed aerial vehicles (UAV) are emerging remote sensing technologies. Unlike satellite- or airborne-based remote sensing (Colpaert, 2022), UAVs being able to fly at lower altitudes, are not susceptible to atmospheric factors, and can obtain data with a high temporal and spatial resolution (Csillik et al., 2018). Therefore, UAVs are widely used in various fields. UAV can be equipped with different sensors for various purposes. Though UAVs equipped with multispectral, hyperspectral, and light detection and ranging (LiDAR) sensors have been in the professional application field for a long time, their high cost and complex data processing process restrict their large-scale promotion and application (Komarek et al., 2022). In contrast, the consumer UAVs equipped with visible-light sensors can provide a potential means for rapid and non-destructive crop information extraction and detection, owing to their low cost, ultra-high resolution, and convenient data processing (Ye et al., 2022).

Recently, high-spatial-resolution remote sensing technology has undergone substantial development (Mehmood et al., 2022; Yang et al., 2022b). High-spatial-resolution images containing large amount of information have also created higher requirements for information extraction and analysis technology. Therefore, to fully exploit the information in high-resolution images, scholars have developed two types of methods based on object-oriented and deep learning (Han et al., 2022b; Liu et al., 2022). Among them, object-based image analysis (OBIA) and machine learning (ML) classifiers have been established as new paradigms in the field of geospatial data and have been applied in several studies (Padua et al., 2022). The object-oriented method uses a set of pixels with characteristics similar to those of the basic units of analysis using a specific scale parameter. It considers the texture, shape, spatial structure, and other multi-dimensional features of adjacent pixels (Filippi et al., 2022). In addition, it combines different ML classifiers, and then conducts information extraction, such as object recognition and classification in a multidimensional feature space (Kamarulzaman et al., 2022). Compared with traditional pixel-based image analysis, OBIA utilizes richer features, improves the interpretability of information extraction results, and facilitates information fusion between remote sensing image processing and geographic information systems (GIS). However, this method has high requirements on the image segmentation scale and selected features, leading to specific defects in its robustness and generalization ability (Siljeg et al., 2022). Existing models often fail to obtain good extraction results in different study areas or with sets of different data. In contrast, deep learning does not require a manual feature design. It can automatically extract the most relevant features of the target task according to the loss function, which has the advantages of solid robustness and easy model transfer (Ghorbanzadeh et al., 2022; Xu et al., 2022). In recent years, target detection and semantic segmentation algorithms in deep learning have made significant progress in crop extraction and counting. Semantic segmentation of images is the process of classifying all pixels in an image and assigning different crop category labels. However, semantic segmentation has non-negligible drawbacks, that is, it can only classify categories and cannot distinguish different individuals of the same category, and it is less effective in processing images with complex information and cannot accurately understand detailed information in images. The target detection algorithm can only identify the number of individuals and their positions and cannot accurately depict the outline of individual plants. Therefore, some scholars have studied the integration of detection and segmentation, and proposed an instance segmentation algorithm. The instance segmentation method combines the advantages of target detection and semantic segmentation ideas to classify all digital image pixels. Based on semantic segmentation, the target detection algorithm is used to distinguish different instances of the same class to achieve instance contour depiction and quantity estimation. He et al. (2017) proposed a two-stage instance segmentation algorithm Mask R-CNN, which is based on Faster R-CNN by adding a few semantics. The algorithm achieved mask prediction and segmentation by adding a small semantic segmentation (FCN) branch to the Faster R-CNN. It proposes an alignment layer to align the extracted features with the input using bilinear interpolation to eliminate the quantization error caused by the RoI pooling in the Faster R-CNN (Ren et al., 2015). Chen et al. (2018) proposed MaskLab to predict the orientation of each pixel concerning its corresponding instance center using orientation features, and then used it to segment instances with the same semantic labels. The aforementioned instance segmentation algorithms are two-stage algorithms, which, like the target detection methods, are highly accurate but slow. Bolya et al. (2019) proposed the YOLACT series of algorithms for one-stage instance segmentation, which segments instances by adding a mask prediction branch to the one-stage detector, thereby achieving fine segmentation of the mask without relying on feature localization. Wang et al. (2020) proposed SOLOv2, a new instance segmentation method that directly detects and segments each instance end-to-end without relying on detection results. Although the deep learning method has shown great application potential in agriculture recently, its advantages and disadvantages in comparison with the current mainstream object-oriented methods have rarely been studied and needs to be explored.

Because of the diversity and irregularity of the spatial morphology of vegetables, an effective extraction scheme has not yet been established, and the existing research is insufficient in feature utilization, sample selection, accuracy verification, and model transferability (Kamga et al., 2021; Wang et al., 2018). Therefore, we propose two methods for vegetable detection: improved object-based image analysis-multilevel distance transform watershed segmentation (OBIA-MDTWS) and deep learning. By comparing the accuracy, stability, and efficiency of the two methods in vegetable extraction in various vegetable planting scenarios, the optimal method for vegetation detection can be determined, providing a new technique for intelligent and informative agricultural production.

2. Materials and methods

2.1. Study area

In this study, a crop field (117°54′E, 26°48′N) planted with cabbage in Shunchang County, Nanping City, Fujian Province, China was selected as the research area. The area has a typical subtropical monsoon climate with an average annual sunshine duration of 1700–1980 h, an average annual precipitation of 900–2100 mm, an average annual temperature of 20–25 °C, an average annual relative humidity of 75–80 %, and a frost-free period of 240–320 days (Ma et al., 2021). This area, which has suitable climatic conditions for crop growth, is a typical artificially planted vegetable field with cabbages of different growth states and plant densities, which is an ideal area for testing the methods used in this study.Fig. 1.

2.2. Framework of study

In this study, we used pixel-based deep learning and OBIA-ML algorithms to detect and count the cabbages based on UAV images, respectively, and the framework of the entire process is shown in Fig. 2: (1) UAV image acquisition with a visible-light sensor; (2) image preprocessing, during which a digital surface model (DSM) and digital orthophoto map (DOM) are generated by Structure from motion (SFM) 3D reconstruction, and the DOM is cropped into sub-maps of uniform size for making cabbage image datasets; (3) cabbage extraction by two different methods, the multi-module method of OBIA and the deep learning method, and (4) accuracy evaluation of the two different methods in cabbage segmentation, and comparison of their accuracy in cabbage extraction and applicability in quantity estimation.

2.3. UAV image data acquisition

A consumer-grade UAV, Mavic Air 2 s (DJI Technology Co., ltd., Shenzhen, China), was used to acquire images within the visible range of

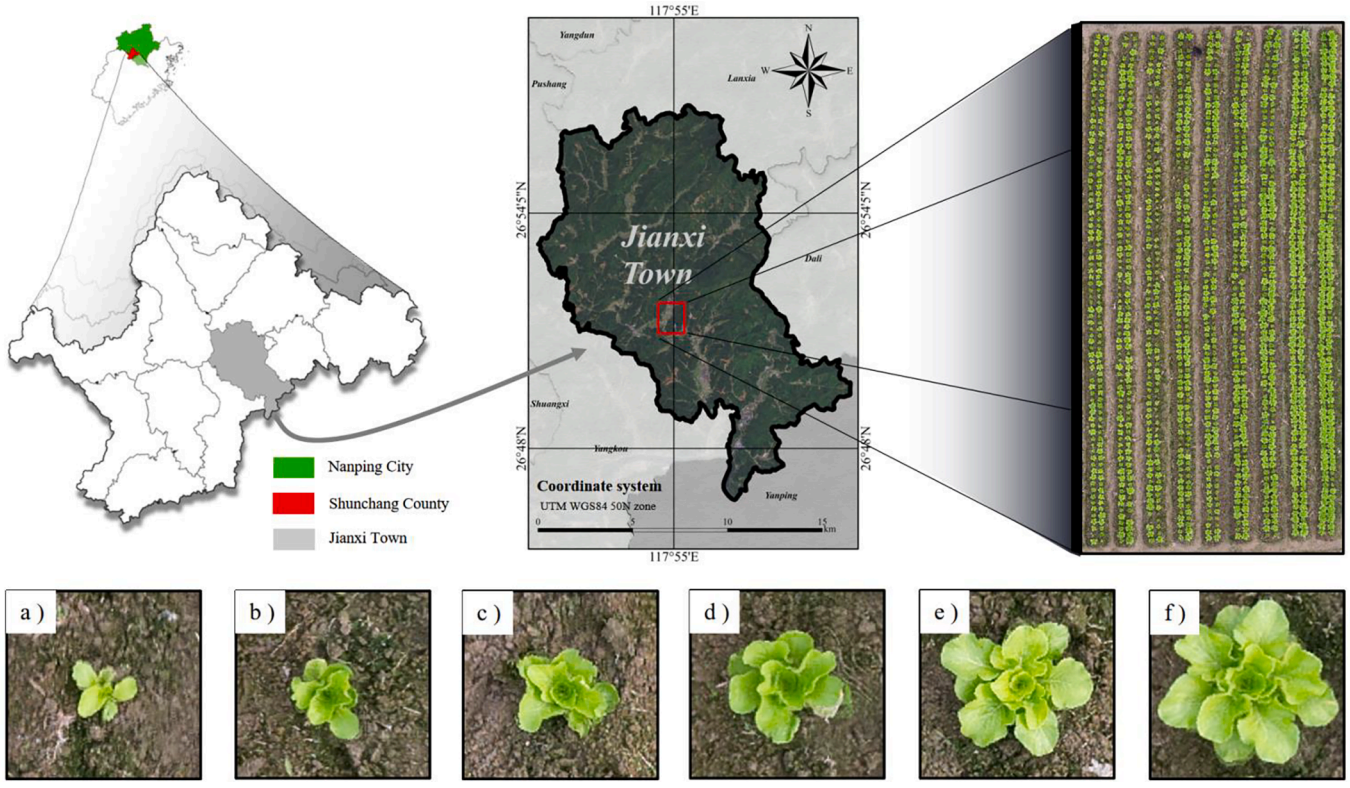


Fig. 1. Geographical location of the study area. (a-f) shows cabbages in different growth stages in the study area.

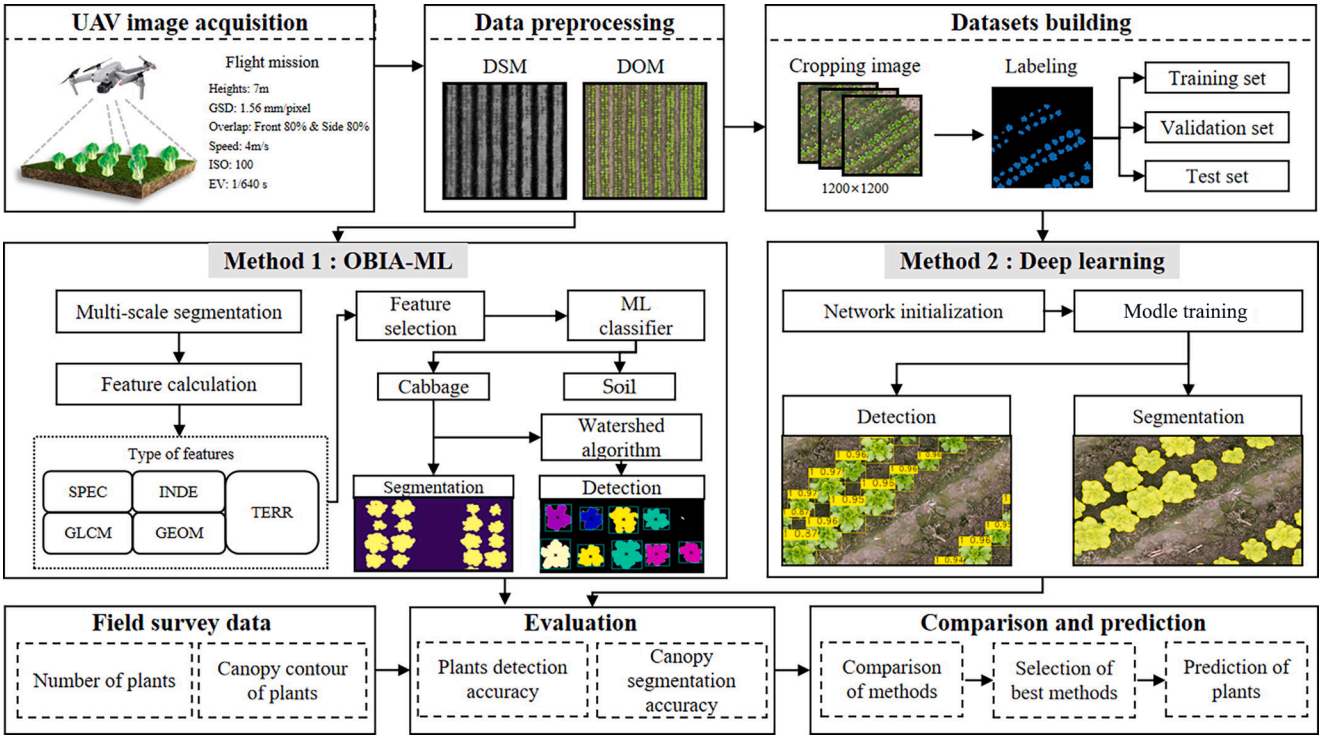


Fig. 2. The framework of cabbage extraction process.

the study area. The UAV has a 1-inch CMOS visible sensor with a pixel size of 2.4 μm, pixel value of 20 million, and camera equivalent focal length of 22 mm. To avoid shadows affecting the detection results, we performed aerial photography on a cloudy day under good lighting conditions. The software named Rainbow 3.8 (Huiteng Software Technology Co., ltd., Zhuhai, China) was used for route planning during the automatic flight mission. The flight altitude was 7 m, the lateral and forward overlap rate were both 80 %, the flight speed was 4 m/s, and the shutter speed and ISO were set to 1/640 s and 100 to avoid blurring of the photos owing to slow shutter speed during the flight. A total of 378

photos were acquired, and an orthophoto with a resolution of 1.56 mm was generated using Pix4D software (Pix4D, Lausanne, Switzerland).

2.4. Deep learning model

According to image segmentation, deep learning can be further divided into semantic and instance segmentation. The semantic segmentation of images is the process of classifying all pixels in an image and assigning different category labels. However, semantic segmentation has non-negligible limitations, that is, it can only classify categories and cannot distinguish between individuals in the same category, it is less effective in processing images with complex information, and it cannot accurately understand detailed information in images. Therefore, to further improve the applicability of semantic segmentation, the instance segmentation method combines the advantages of target detection and semantic segmentation to classify all digital image pixels and distinguish different instances of the same category based on semantic segmentation to achieve instance contour depiction and quantity estimation.

2.4.1. Dataset production and data enhancement

Owing to the large size of the orthophoto and the limitation of computer performance, the orthophoto cannot be completely input into the network at one time; therefore, it is necessary to slice it to ensure network training and detection speed (Shi et al., 2021). In this study, the orthophoto was divided into sub-images of size 1200 × 1200 pixels. The software named Labelme 3.6 was used to label the data according to the real edges of the cabbage, and the cabbage dataset was constructed in VOC format. The sample dataset was divided into three parts, with 70 % used for model training (140 images, total 987 plants), 20 % used for model validation during training (40 images, total 282 plants), and the remaining 10 % (20 images, total 141 plants) used to evaluate the model performance.

Deep learning is a data-driven technique, the model of which usually has a complex network structure (Qin and Liu, 2022). Therefore, extensive training samples are required to enable the deep learning model to thoroughly learn the multi-level features of the target object and avoid being affected by the complex background of the original image and uneven light intensity. However, the area of artificially planted crops is often small, the image data that can be used as training samples are limited, and the deep learning model has high requirements for training samples; therefore, it is prone to overfitting in small-scale samples (Paoletti et al., 2022) (Fig. 3). Hence, this study used data augmentation techniques to expand the sample set to solve the overfitting problem; thus, the model can obtain the best performance and high robustness. This study used geometric and spectral transforms to expand the original dataset for data augmentation. The geometric transformation adopts random flip, random rotation (0°, 90°, 180°, 270°), random scaling (approximately 20–200 %), and random cropping to transform the image's geometry, enhancing the model's adaptability to different shooting angles, flying heights, and sizes of individual vegetables. Spectral transformation transforms the image in the HSV color space, such as contrast (42–56), hue (86–92), saturation (82–109), and brightness (119–138), so that the trained model can fully learn the characteristics of vegetables under different light conditions and growth states. Finally, the 200 training images were expanded to 1600 images through image enhancement.

2.4.2. Two stage instance segmentation algorithm

Mask R-CNN algorithm consists of two main branches (Han et al., 2022a; Wu et al., 2021): detection and segmentation (Fig. 4). The detection branch realizes the localization and classification of the target in the image, and the segmentation branch generates a binary mask through the fully convolutional network (FCN) to realize contour segmentation of the crop and the pixel-level extraction effect. The overall framework of the model employed in this study is illustrated in Fig. 4. First, the image to be detected is passed into the model, and the features of the entire image are extracted through a convolutional neural network to obtain the corresponding feature map. Subsequently, in the classification branch, the target frame is located and classified. The corresponding binary mask is drawn on the image in the segmentation branch using a fully convolutional network to realize instance segmentation. Finally, the predicted image with the bounding box and outline of the cabbage contour is output.

The MaskLab algorithm is a two-stage instance segmentation algorithm based on the improved Faster R-CNN target detection algorithm with three output branches: target bounding box, semantic segmentation, and direction prediction (Fig. 5). The algorithm first uses the Faster R-CNN algorithm to provide instance target frames with accurate location information. Then, within each region of interest, the corresponding semantic channels are selected and cropped according to the corresponding class. At the same time, the regional logits from each channel are combined using direction pooling. Finally, these features are convolved by 1 × 1 to obtain the object contour mask.

2.4.3. Single stage instance segmentation algorithm

The YOLACT network consists of a feature extraction network and image function network (Fig. 6). The feature extraction network module mainly performs image feature extraction through the target detection network (RetinaNet), which consists of a residual network (ResNet-101)

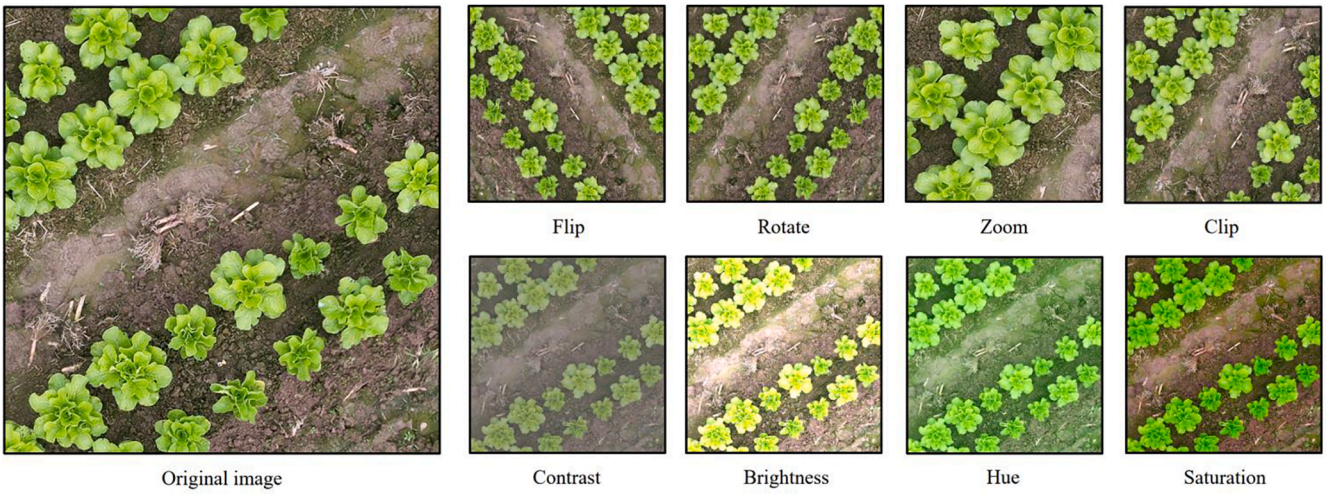


Fig. 3. Data augmentation.

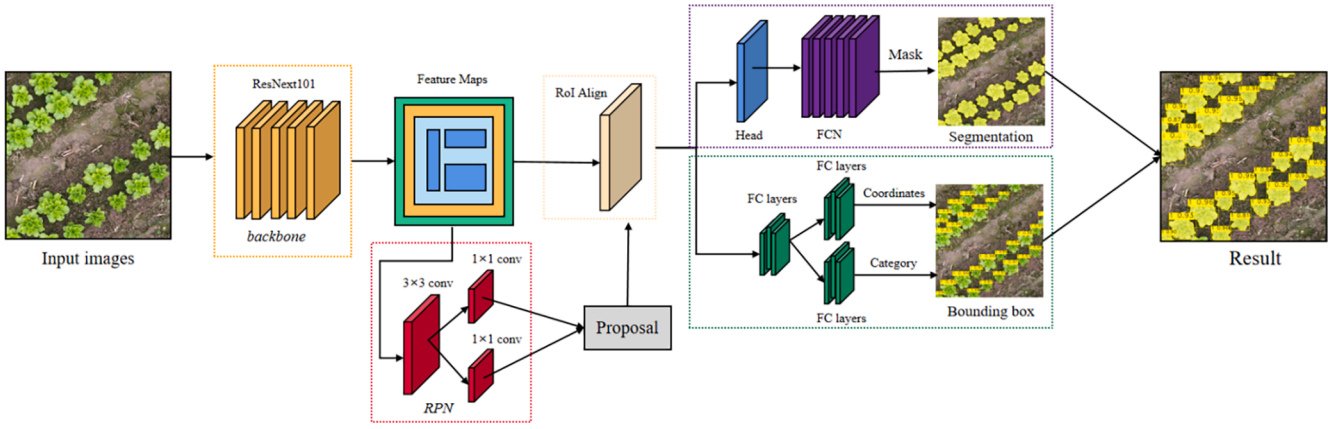


Fig. 4. The structure diagram of Mask R-CNN network.

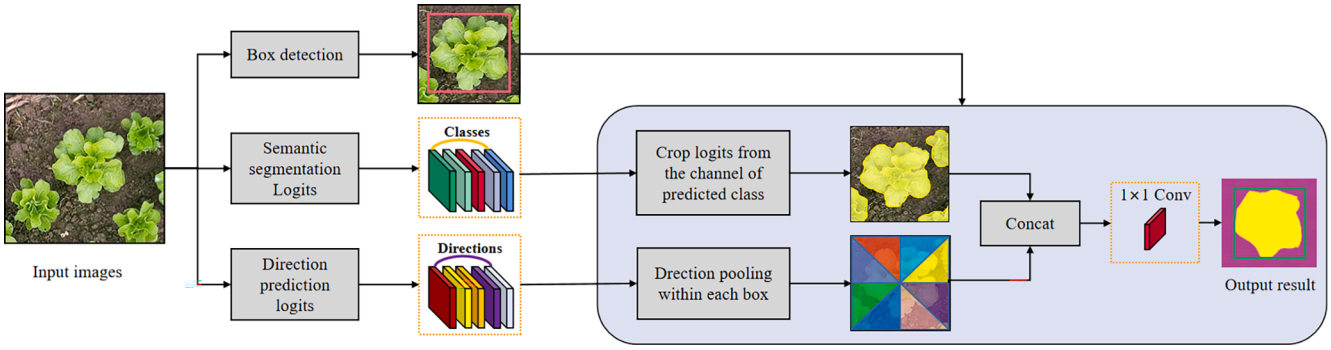


Fig. 5. The structure diagram of MaskLab network.

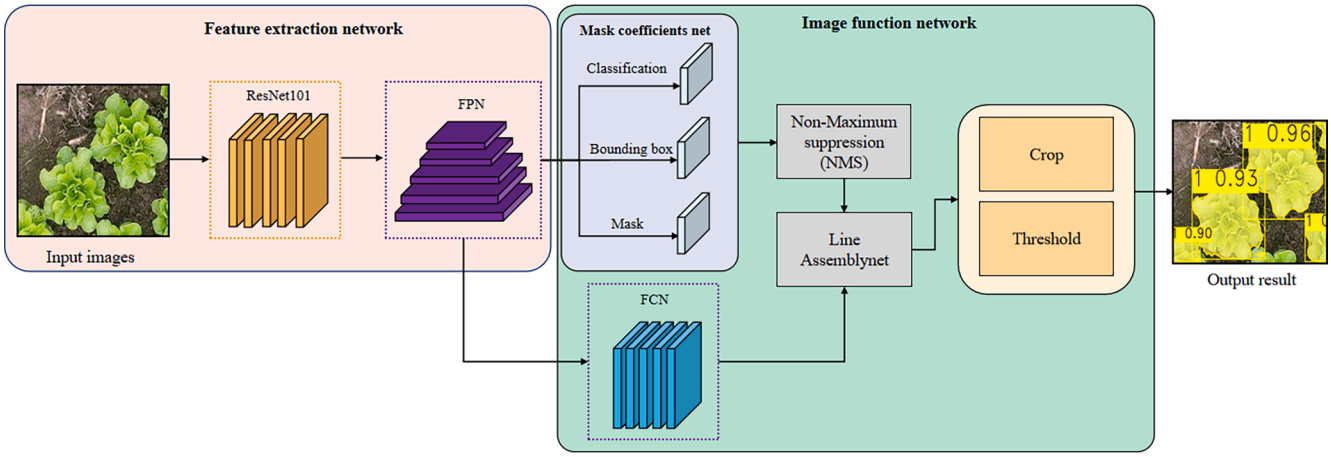


Fig. 6. The structure diagram of YOLACT network.

and a feature pyramid network (FPN). The image function network module is divided into two branches. One branch has the main task of completing the prediction of the mask coefficients and obtaining the three results of the classification probability of the target, the probability of the bounding box, and the mask prediction probability, and then obtaining the prediction result with the maximum confidence through the non-maximum suppression (NMS) algorithm. The other branch completes the prototype mask through a full convolutional neural network (FCN). The two branch tasks are parallel tasks, computed separately and in parallel, to significantly improve the computation speed. After the completion of two branching tasks, the optimal segmentation map of each instance is obtained using a linear combination. Finally, the optimal segmentation map is cropped and adjusted by a threshold value for each instance to obtain the final intuitive image segmentation effect.

SOLOv2 inherits the model architecture of SOLO (Fig. 7), divides the input image into S × S grids, and predicts the semantic category and instance mask of the object by the category branch and the mask branch that is connected to the FPN pyramid layer after the center of the object. This breaks through the traditional local border detection and pixel point aggregation methods and achieves end-to-end optimization, and is a better performing instance segmentation network. After image input, the full convolutional network (FCN) is responsible for generating feature images with different scales and high-level semantic

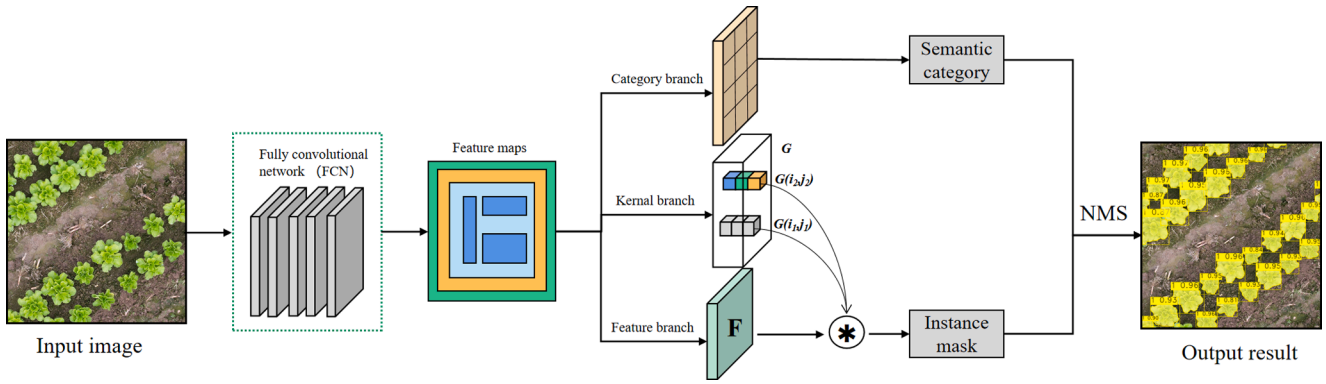


Fig. 7. The structure diagram of SOLOv2 network. Note: G is the dynamic convolution kernel matrix generated by the kernel branch, G(i_1,j_1) and G(i_2,j_2) represent the convolution kernels at grids (i_1,j_1) and (i_2,j_2) respectively, F is the high-resolution mask feature matrix generated by the feature branch, and the matrix NMS represents the matrix non-maximum suppression.

information, whereas the category branch performs category prediction for each grid in the feature image and classifies a target center as a positive instance when it is within the grid and satisfies the threshold condition. Meanwhile, the product of the dynamic convolution kernel G generated by the kernel branch, feature branch, and high-resolution mask feature matrix F is used in the mask branch to generate the corresponding category mask and solve the problems of a large number of parameters and redundancy of prediction results in mask prediction. Finally, the matrix NMS technique is used to significantly reduce the inference operation overhead of the algorithm to further improve the detection rate. Its loss function consists of semantic category loss and mask prediction loss.

2.4.4. Deep learning model training

The hardware and software environments used in this experiment, are listed in Table 1. This study used pre-trained weights from the MS-COCO dataset for this task, and the constructed small-sample cabbage dataset was trained using migration learning. This method accelerates the convergence of the model and the computational cost associated with model training. The training parameters were set as follows: the batch size was set to 4, the number of iterations (epochs) was set to 300, and the NMS threshold was set to 0.2. Because the model weights are randomly initialized at the beginning of training, selecting a larger learning rate at this time may cause instability (oscillation) in the model. Therefore, we selected the learning rate warm-up, which can reduce the learning rate in several epochs or certain steps at the beginning of training. Under a low learning rate of preheating, the model gradually becomes stable. After the model became relatively stable, the preset learning rate was selected for training, which made the model converge faster and improved its performance. Therefore, in this study, the initial learning rate was set to 0.00125 and the threshold of the warmup step was set to 400.

2.5. OBIA method

2.5.1. Image segmentation

OBIA technology divides an entire image into smaller image objects with similar features through a segmentation procedure. Therefore, it can use more features, such as spectral, shape, and texture, than pixel-based methods (Guo et al., 2022; Li et al., 2022). It has been widely proven that OBIA technology can achieve a higher classification accuracy in various fields than pixel-based methods.

Image segmentation is a critical step in object-oriented image processing, and the segmentation quality significantly impacts the extraction accuracy of object-oriented algorithms (Gao et al., 2022). In this study, we used six algorithms for image segmentation of OBIA, including (1) multi-resolution segmentation from proprietary software (Trimble eCognition); (2) mean shift from proprietary software (Esri's ArcGIS); (3) regional growth from geographic resources analysis support system (GRASS GIS) software; (4) large-scale mean shift from orfeo toolbox (OTB); (5) shepherd k-means iterative elimination from the remote sensing and gis software library (RSGISLib); (6) mean region growing from system for automated geoscientific analyses (SAGA GIS). Multi-resolution segmentation, as a common segmentation algorithm, merges the pixels into a larger clustered and homogeneous patch from the bottom up based on the homogeneity criterion. The segmentation scale defines the heterogeneity threshold at which the growth of this object stops. The segmentation object heterogeneity H includes spectral heterogeneity h_{color} and the shape heterogeneity h_{shape} (Eq. (1)). After repeated tests, the spectral weight and tightness factor weight were set to 0.9 and 0.5, respectively, and the weight for each layer was equal in this study.

$$\begin{cases} H = \omega_{color} h_{color} + (1 - \omega_{color}) h_{shape} \\ h_{shape} = \omega_{compact} h_{compact} + (1 - \omega_{\mathrm{compact}}) \mathrm{h}_{\mathrm{smooth}} \end{cases} \quad (1)$$

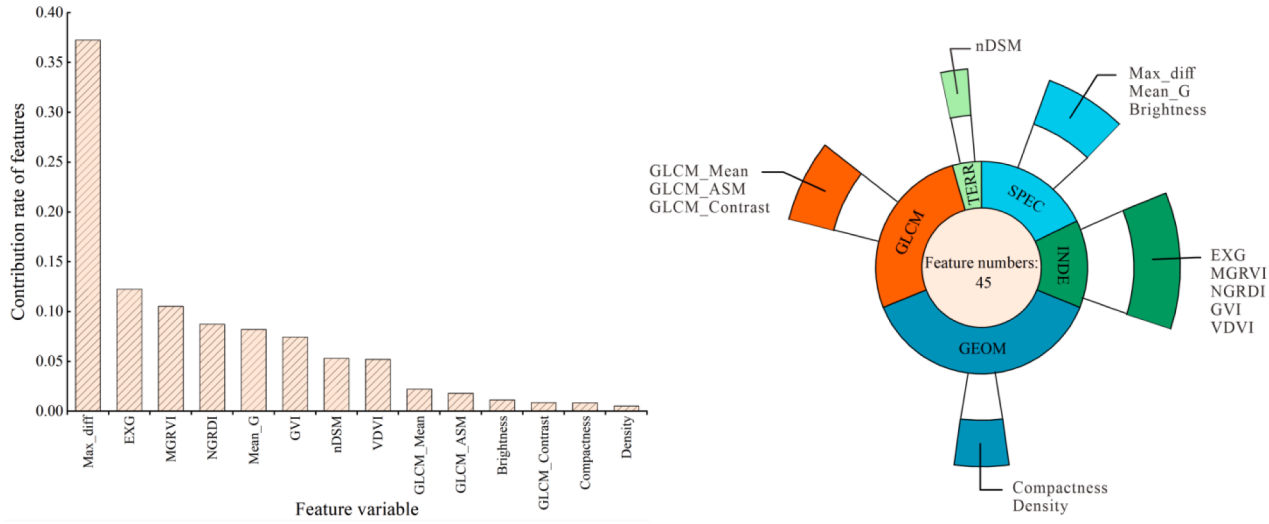
Where ω_{color} is the spectral weight, (1-ω_{color}) is the shape weight, h_{shape} is the shape heterogeneity consisting of tightness, $h_{compact}$, smoothness h_{smooth}, $\omega_{compact}$ denotes the tightness factor weight, and (1-$\omega_{compact}$) represents the smoothness weight factor.

Table 1
Software and hardware specifications.

Name	Parameters and versions
CPU	Intel Core™ i7-12700 K @4.9 GHz
RAM	64 GB (Kingston DDR4 3200 MHz)
SSD	HS-SSD-C2000Pro 2 TB
GPU	NVIDIA Tesla V100 (32 GB)
OS	Windows 10 Professional (DirectX 12)
ENVS	PyTorch 1.9.0 + Python 3.8 + R 4.2.2

2.5.2. Extraction of feature

Yang et al. (2022a) pointed out that classification accuracy is highly correlated with the number of features. Therefore, we selected five types of features of the objects: spectral, geometric, texture, terrain, and vegetation indices. To avoid noise or redundant information brought by multi-dimensional features and improve the computational efficiency and recognition accuracy, we used a recursive feature elimination (RFE) algorithm to reduce the feature dimensionality. RFE is a greedy optimization algorithm that can obtain the best feature subset by eliminating the feature with the lowest contribution rate in each iteration. We implemented RFE using the sklearn package in Python. In addition, we determined the importance of each feature through the cross-validation of the tree nodes (Fig. 8). 14 effective features were retained after feature screening (Table 2).



a. The importance of optimal features b. Proportion of optimal features of different types

Fig. 8. Optimal features through a recursive feature elimination (RFE) algorithm.

Table 2
Selected features.

Feature type	Feature name	Formula	Feature descriptions	References
Spectral feature	Mean_G	$\frac{1}{N}\sum_{i=1}^{N} G_i$	The mean value of the object in the green band	(Zhang et al., 2022a)
	Brightness	$\frac{1}{3N}\sum_{i=1}^{N} R_i + G_i + B_i$	The sum of the average gray-scale values of all pixels within the object	(Feng et al., 2022)
	Max_diff	$\frac{\frac{1}{N}\sum_{i=1}^{N}(Max(R,G,B) - Min(R,G,B))}{\text{Brightness}}$	Maximum difference in gray-scale values between different bands	(Gurunathan and Krishnan, 2022)
Geometric feature	Compactness	–	Extent to which the object is filled with its outer rectangle	(Pendyala et al., 2021)
	Density	–	Inversely proportional to the proximity of the object to the rectangle	(Pendyala et al., 2021)
Vegetation index	EXG	$2\times G - R - B$	Widely used in the task of distinguishing weeds from non-plant backgrounds	(Zhang et al., 2022b)
	GVI	$G/(R+G+B)$	Widely used in grassland degradation monitoring, quantitative analysis of grassland resources, etc.	(Feizi et al., 2022)
	MGRVI	$(G^2 - R^2)/(G^2 + R^2)$	Widely used in biomass and yield estimation research	(Feng et al., 2022)
	NGRDI	$(G-R)/(G+R)$	Widely used in growth monitoring and crop identification tasks	(Cobelo et al.)
	VDVI	$(2\times G - R - B)/(2\times G + R + B)$	Widely used in green space information classification and vegetation coverage estimation tasks	(Zhang et al., 2022a)
Texture feature	Mean	$\sum_{i=1}^{N}\sum_{j=1}^{N} i\, P_{i,j}$	Represents the average gray-scale of the object	(Gurunathan and Krishnan, 2022)
	Contrast	$\sum_{i=1}^{N}\sum_{j=1}^{N}(i-j)^2 P_{i,j}$	A measure of the local gray-scale difference of the object	(Gurunathan and Krishnan, 2022)
	Angular second moment	$\sum_{i=1}^{N}\sum_{j=1}^{N} i\, P_{i,j}^2$	A measure of the uniformity of the object gray distribution and the thickness of the texture	(Pantic et al., 2015)
Terrain feature	Normalized digital surface model (nDSM)	*DSM-DTM*	A model that can reflect the real height of the ground object is obtained through the difference operation between the DSM and the digital terrain model (DTM)	(Pendyala et al., 2021)

Note: there is no formula for spectral and geometric features, for they are the inherent properties of the object.

2.5.3. ML classification algorithms

In recent years, ML algorithms have been widely used for optical remote sensing image classification tasks. Commonly used supervised ML classifiers include the k-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF). KNN is a non-parametric machine-learning algorithm. It finds the nearest training sample for the test sample through the distance measure and predicts according to the category of the training sample. Compared with other algorithms, the KNN is relatively easy to implement. An SVM is a binary classification model that aims to find a hyperplane in a high-dimensional feature space to separate training samples, which can solve small sample size, nonlinear, and high-dimensional binary classification problems. Among these algorithms, RF accounts for a considerable proportion. The RF is a nonlinear and nonparametric ML algorithm (Li et al., 2021). It integrates multiple classification and regression tree (CART) through ensemble learning, uses multiple trees to train and predict samples, and uses voting to determine the classification results of samples. The RF model has a strong analytical power and is robust when dealing with high-dimensional features and complex datasets. However, similar to most ML algorithms, they also have the computational difficulty of dealing with high-dimensional data and require a large amount of training data to tune the structural parameters.

The value of k is the only hyper-parameter of the KNN model, and the meaning of k is the number of reference "neighbor" label values. When the value of k is small, the model complexity is high, the training error is reduced, and the generalization ability is weakened; when the value of k

is large, the model complexity is low, the training error is increased, and the generalization ability is somewhat improved. The SVM model has three very important parameters, kernel, C and gamma, where C is the penalty factor, i.e., the tolerance for error. the higher the c, the more intolerant the error is and the easier it is to over fit. the smaller the c, the easier it is to under fit. gamma is a parameter that comes with the radial based Function (RBF) function after choosing it as the kernel. The larger the gamma value, the less support vectors, and the smaller the gamma value, the more support vectors. The two most important parameters of the random forest model are the number of decision trees and the number of features. Usually, the larger the number of decision trees, the better the model is, but the accuracy tends to stabilize when a certain number is reached. The accuracy of the model is highest when the RF model has the optimal number of features.

We used RFE method to optimize the SVM and RF model parameters. The ML model parameters used in this study are shown in Table 3, where we set the kernel of the SVM model to RBF, C to 4, and gamma to 0.02; the number of decision trees in the RF model is 378 and the number of features is 45; and the K value of the KNN model is set to 5.

2.5.4. Segmentation parameter optimization

Parameter optimization, particularly automated parameter-tuning algorithms, has become a recent research hotspot. Generally, these algorithms complete parameter selection by combining all parameters or by fixing other parameters and tuning only one parameter simultaneously. These algorithms require several trials and are time-consuming, which is difficult to achieve in multi-parameter and multi-level experiments. Therefore, we used an R package called SegOptim (Goncalves et al., 2019), which integrates the segmentation and classification steps of OBIA into a complete system and optimizes each parameter of this system using a genetic algorithm (GA). GA is a method of searching for the optimal solution by simulating the natural evolution process. The algorithm uses a global random search strategy, and through operations such as selection, inheritance, and cross-mutation, it improves the fitness of individuals participating in evolution, and then optimizes and solves complex problems. To run the GA algorithm stably, several main parameters must be set and determined through repeated experiments (Table 4).

Although this open-source tool does not support parameter optimization in the proprietary software eCognition, we used an automated tool integrated with a plug-in (ESP2) to select the scale parameters for multi-resolution segmentation. The estimation of scale parameter (ESP) is a bottom-up region-merging technique that can generate optimal segmentation parameters by iterative segmentation. Determination of the optimal segmentation scale is usually expressed by the local variance (LV) of the homogeneity of the segmented object and its rate of change (ROC, Eq. (2)). With an increase in the segmentation scale, multiple local maximum points appeared in the ROC curve, and each local maximum point may be the best segmentation scale. To maximize the homogeneity of the object, the first local maximum point of the ROC curve is chosen as the best segmentation scale.

$$ROC = \frac{L_i - L_{i-1}}{L_{i-1}} \times 100\% \quad (2)$$

where ROC is the rate of change of LV, L_i is the average standard deviation of the i th object layer, and L_{i-1} is the average standard deviation of the i-1 th object layer.

Table 3
ML model parameter.

ML model	Model parameter	Optimization method
k-Nearest neighbors (KNN)	K = 5	–
Support vector machine (SVM)	Kernel = RFB, C = 1, Gamma = 0.02	SVM-RFE
Random forest (RF)	Decision trees = 378, Features = 14	RF-RFE

Table 4
Genetic algorithm (GA) parameter details.

GA parameters	Feature descriptions	Value
Initial population	The population size of the GA after initialization	50
Max iteration	Controlling the maximum number of times a GA run ends	100
Mutation probability	The probability that a partial random parameter in a subpopulation produces a random value	0.05
Crossover probability	Controlling the frequency of the population crossover operation used	0.5

2.5.5. Morphological image postprocessing

The object-oriented image processing method can eliminate salt-and-pepper noise in the image and realize high-precision extraction of cabbage. However, the extracted image still contained holes, burrs, and misclassification. Therefore, we used the mathematical morphology method to preprocess the images. The workflow is illustrated in Fig. 9. First, we used median filtering to process the image because the watershed segmentation algorithm based on grayscale used to find the segmentation line, is susceptible to noise in the image. The advantages of median filtering are simple calculation process, good filtering effect on impulse noise, along with the complete edge information of the image (Evagorou et al., 2022). Second, we used a flooding algorithm to deal with the internal holes caused by the segmentation of uneven gray areas so that the extracted lettuce maintains internal continuity and smooth edges, which is convenient for subsequent segmentation operations.

2.5.6. MDTWS individual plant segmentation

We used a topology-based watershed segmentation algorithm to count the number of individual cabbage plants with high density and adhesion. The topology-based watershed algorithm is an essential morphological segmentation method that is widely used in image segmentation because of its positive effect on region edge localization and closed contour extraction (Xue et al., 2021). The watershed algorithm is sensitive to weak edges. It does not easily lose edge information, which is the key to solving the adhesion problem. However, it also causes image oversegmentation, affecting segmentation accuracy. Therefore, we used a watershed segmentation method combining morphology and Euclidean distance transformation to perform distance calculations (Eq. (4)) from each pixel in the image to the nearest non-zero-valued pixel. The binary image of the adhesion region can form more evident valley lines after the distance transformation. These valley lines can be marked as segmentation lines between adhesion particles. However, single-level distance transformation can easily cause under-segmentation or over-segmentation of the target object, and it is not suitable for separating severely adhered particles (Okorie and Makrogiannis, 2019). Considering this issue, we propose a watershed segmentation algorithm based on a multi-level distance transformation (Fig. 10). First, the normalized image was traversed to extract the target contour area, and the average area of all contours was taken as the threshold. Targets with an area smaller than the threshold were retained, and targets with a contour area larger than the threshold were subjected to secondary distance transformation. Several iterations were performed until the adhering targets were well separated. Second, the image after multi-level distance transformation was used as the marker, and the watershed algorithm was used for calculation. The algorithm can avoid over- and under-segmentation and ensure the integrity of the edge information of the adherent particles to the maximum extent.

$$Dist_{ij} = \min[\sqrt{(i-\mathrm{x})^2 + (\mathrm{j}-\mathrm{y})^2}] \ ((x,y) \in P) \quad (3)$$

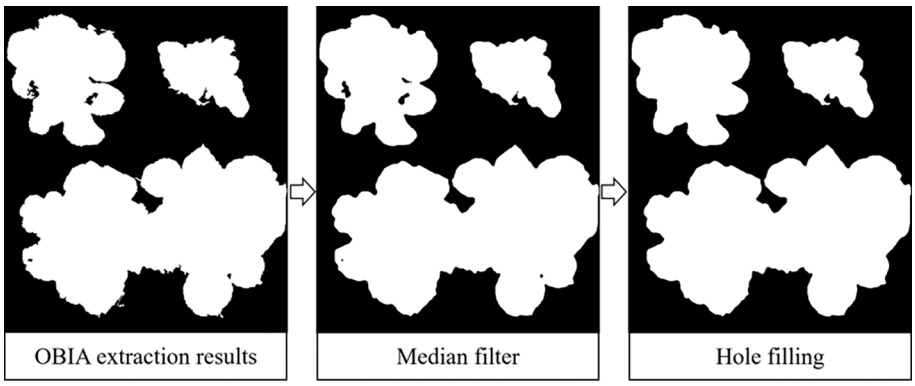


Fig. 9. Morphological processing.

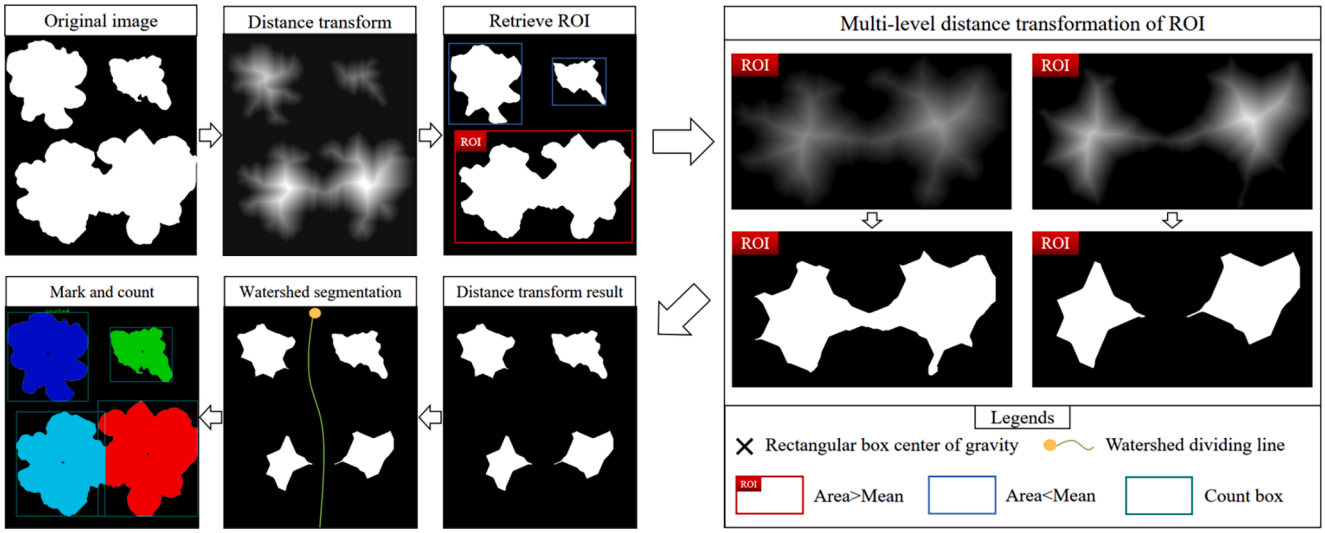


Fig. 10. Diagram of distance-based transformed watershed algorithm.

where $Dist_{ij}$ is the shortest distance from pixel (i,j) to the nearest non-zero valued pixel (x,y).

2.6. Accuracy evaluation

For target extraction, the model performance is often evaluated by comparing the number of pixels between the true category of the validation sample and the predicted category of the model. There are generally-four cases: True Positive (TP), that is, the extraction result is consistent with the true category; False Positive (FP), that is, the actual situation is the background but is mistakenly identified as the target crop; False Negative (FN), that is, the crop in the real scene is not correctly detected, and True Negative (TN) is the pixel that is correctly identified as the background. According to these four scenarios, this study used precision (Eq. (4)) and recall (Eq. (5)) to evaluate the accuracy of the two cabbage extraction methods. In the process of training models, it is often necessary to balance the precision and recall, and the F1-Score is the summed average of precision and recall, which can well unify precision and recall. Therefore, we used the F1-Score (Eq. (6)) as the overall evaluation index for extraction performance in this study. According to these four situations, this study used the miss detection rate, false detection rate, and accuracy indicator to evaluate (Eq. (7)–(9)), which were calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

$$Miss\ detection\ rate = \left(\sum \frac{y_{mi}}{y_i}/n\right) \times 100\% \tag{7}$$

$$False\ detection\ rate = \left(\sum \frac{y_{fi}}{y_i}/n\right) \times 100\% \tag{8}$$

$$Accuracy\ rate = \left(1 - \left|\frac{\hat{y} - y_i}{y_i}\right|\right) \times 100\% \tag{9}$$

Note: y_i is the number of true cabbages grown in the study area, y_{mi} is the number of unidentified cabbages, and y_{fi} is the number of misidentified cabbages.

3. Results

3.1. Optimal solution selection for OBIA and deep learning

To select representative models in the OBIA and DL methods, we used several mainstream deep learning models capable of individual detection and combined multiple segmentation algorithms and ML algorithms in OBIA. We accurately depicted the real contour of each object by zooming in on the orthophoto and compared it with the predicted

results of the two schemes to ensure accuracy of the results. For the OBIA, we used six different segmentation algorithms and three classifiers. The parameters of the segmentation method or classifier were optimized using the GA algorithm integrated in the SegOptim package. For multi-resolution segmentation from eCognition 10.1, we employed the proprietary tool ESP2 to optimize its parameters. The results showed that the OBIA method based on the parameter optimization scheme achieves good results in terms of accuracy. For the entire image, the overall accuracy values range between 83.79 and 96.20 %. Among them, the RF and SVM methods performed the best (out of the six segmentation schemes, RF and SVM were selected as the best classifiers three and two times, respectively). Although several segmentation schemes such as multi-resolution, mean-shift, and shepherd iterative achieved similar results (refer Table 5), the data showed that multi-resolution segmentation yields better results for two of the three classifiers. Among all schemes, the combination of multi-resolution and RF attained the optimal results (96.20 % accuracy).

For DL, instance segmentation algorithms can distinguish between adjacent objects of the same semantic class and realize individual detection. We used PA and mAP to measure and evaluate the performance of mainstream instance segmentation networks for image segmentation and object detection tasks, respectively. As can be observed from the results (Table 6), the Mask R-CNN network that we used achieved the best results in the evaluation of segmentation accuracy and target detection accuracy (PA = 96.50 %, mAP = 86.63 %). However, it is not the best network in terms of the inference speed. Although we are interested in the inference time of the network, for our limited dataset, it is quite cost-effective to sacrifice inference time for the improvement of the model prediction accuracy.

3.2. Analysis of plant extraction results

To better compare the performances of pixel-based deep learning and OBIA-ML algorithms, we selected the best algorithms (Mask R-CNN and OBIA-MDTWS) for DL and OBIA. We then tested these in four representative regions (A–D) with different plant densities and growth stages of cabbages. It can be observed that both OBIA-MDTWS and the Mask R-CNN can separate cabbages from soil relatively completely (Fig. 11), the edges are clear, and the extraction effect is good. However, the OBIA-MDTWS method misclassifies a small amount of straw distributed in the soil as cabbage. In addition, the OBIA-MDTWS method also misclassifies the marginal soil as cabbage leaves, the edge information of the cabbage individual mask images is missing, and there are holes inside. Overall, the extraction results are relatively fragmented, which may be related to the OBIA segmentation scale. In contrast, the Mask R-CNN has no misclassification or hole phenomena. In regions with medium plant densities and large individuals (C–D), OBIA-MDTWS misclassified some weeds as cabbages, especially in region D. In contrast, the Mask R-CNN deep learning model has a relatively better extraction effect, and there is no case of classifying weeds as cabbages. There is a high degree of agreement between the segmented and actual edges, and individual instances can be better distinguished. Overall, the extraction segmentation result of Mask R-CNN was better than that of OBIA-MDTWS.

To verify the performance of the Mask R-CNN and OBIA-MDTWS methods, we conducted an accuracy evaluation and analysis of the above four regions based on the confusion matrix. It can be noted that the precision of OBIA-MDTWS method for cabbage extraction in different regions (A–D) are 93.81, 94.19, 93.23 and 92.25 %, respectively, and the recall rates are 96.72, 97.08, 96.36, and 96.08 %, respectively (Fig. 12). The average F1-Score value of the four regions was 94.93 %, each higher than 90 %, indicating a good overall performance. The performance of the Mask R-CNN deep learning model was better compared with that of OBIA-MDTWS. The precision of cabbage extraction in four regions was higher than 96 %, with the lowest value of 96.14 % and the highest value of 97.62 %, and the overall misclassification phenomenon was less. The lowest recall rate was 98.01 %, and the highest value was 98.61 %, the lowest F1-Score was 97.36 %, and the highest value was 97.81 %. The average F1-Score value was 97.63 % for the Mask R-CNN, which was 2.70 percentage points higher than that of OBIA-MDTWS. This suggests that the Mask R- CNN method performs better than the OBIA-MDTWS method for extracting cabbages.

3.3. Analysis of cabbage counting results

Fig. 13 shows the results of individual segmentation and counting of cabbage in multiple regions of the study area with different plant densities and growth states, using the OBIA-MDTWS and Mask R-CNN models. Because Mask R-CNN provides an end-to-end learning paradigm, and the output results map the original data, it can visualize the

Table 5
Accuracy comparison of different OBIA segmentation algorithms combined with different classifiers.

Software	Segmentation algorithm	Accuracy (%)			Optimizer
		SVM	KNN	RF	
eCognition	Multi-resolution	95.99	90.28	**96.20**	ESP2 + GA
ESRI® ArcGIS	Mean-shift	92.21	84.39	**93.82**	GA
GRASS GIS	Region growing	**89.73**	86.89	90.44	GA
SAGA GIS	Mean region growing	90.97	**90.65**	91.35	GA
Orfeo Toolbox	Large scale mean-shift	87.29	84.11	**86.87**	GA
RSGISLib	Shepherd iterative	**92.17**	83.79	89.55	GA

Note: Bold values means the best value among combination of multiple classification methods.

Table 6
Comparison of different deep learning instance segmentation algorithms.

Deep learning algorithm	Backbone	Segmentation accuracy	Object detection accuracy	Time (s)
		PA(%)	mAP(%)	
Mask R-CNN	ResNeXt-101-FPN	**96.50**	86.63	1.04
SOLOV2	ResNet-101-FPN	92.28	85.14	**0.38**
MaskLab	ResNet-101	92.86	82.75	2.12
YOLACT	ResNet-101-FPN	91.18	77.45	0.67

Note: Bold values means the best value among combination of multiple classification methods.

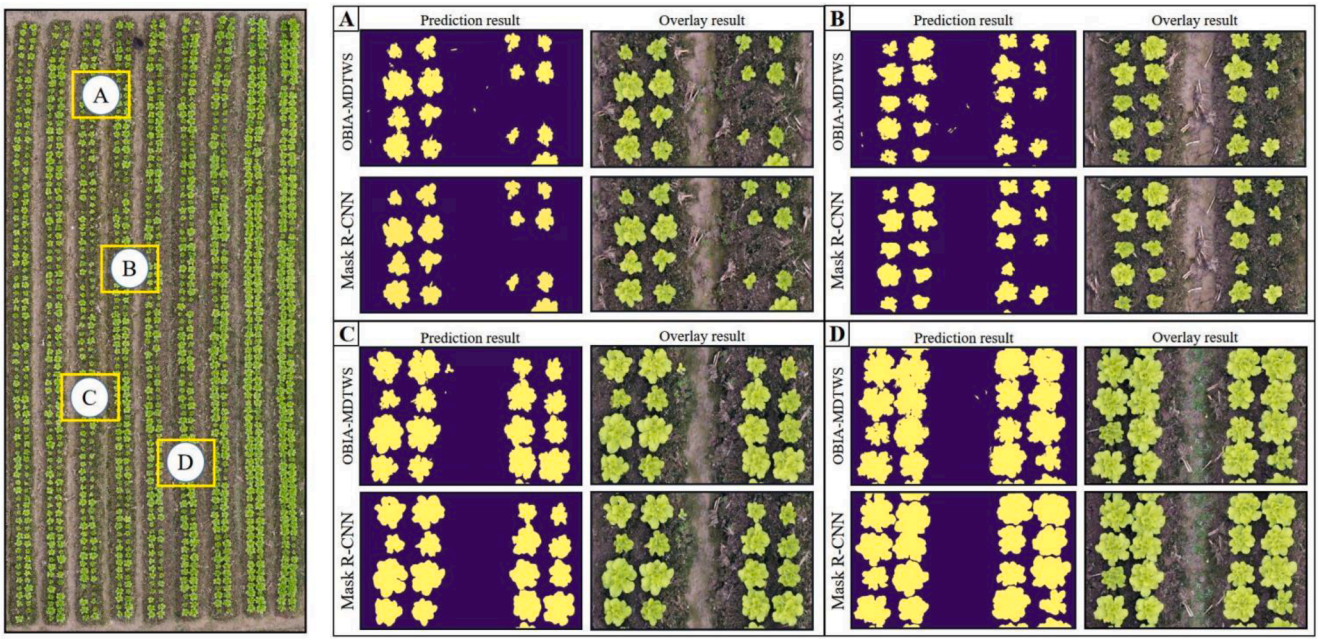


Fig. 11. Extraction results of OBIA-MDTWS and Mask R-CNN methods.

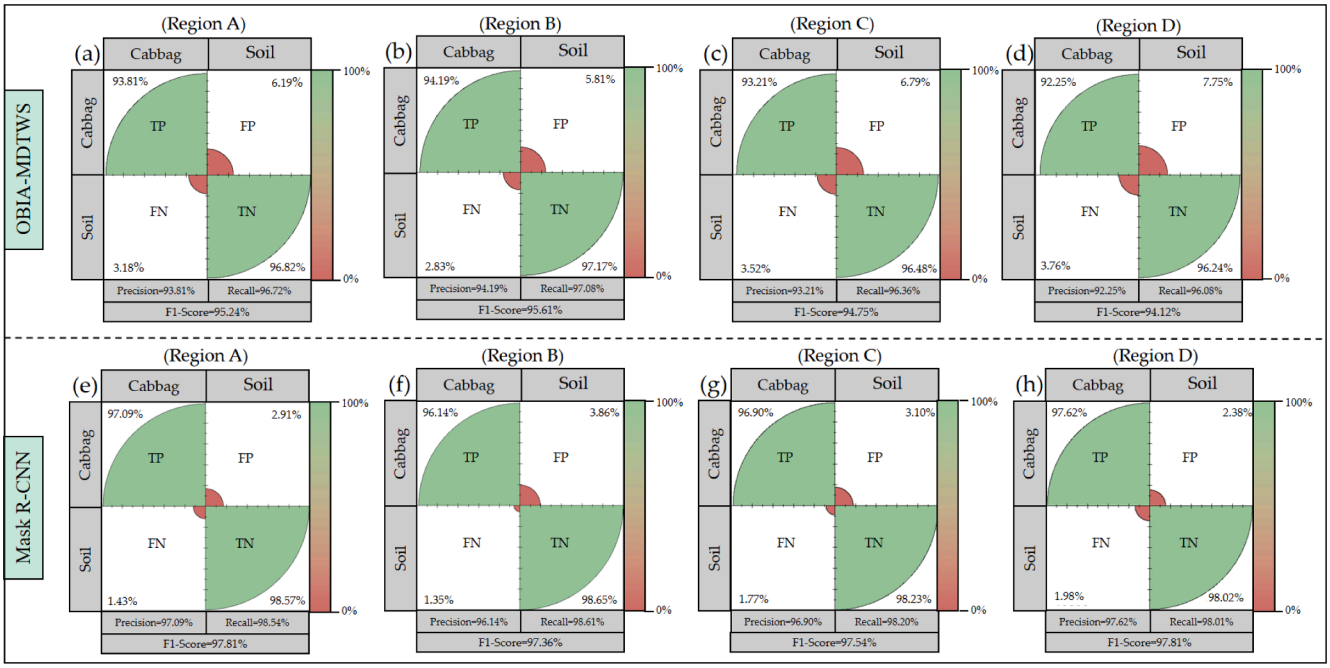


Fig. 12. Extraction accuracy of OBIA-MDTWS and Mask R-CNN methods.

localization effect of different cabbage individuals in the original images. For the output results of the OBIA-MDTWS method, we used a mask of different colors to distinguish cabbage individuals, external rectangular boxes, and the contour center of gravity for individual cabbage localization. Overall, the OBIA-MDTWS and Mask R-CNN methods showed good segmentation results for cabbage individuals with different degrees of adhesion. The Mask R-CNN deep learning model was better than the OBIA-MDTWS model for the detection of cabbage individuals. The OBIA-MDTWS model proposed in this study effectively suppresses the under-segmentation phenomenon of traditional watershed algorithms. However, owing to the influence of noise and local fractures generated by multiple distance transformations, oversegmentation occurs for individuals with complex edges or irregular contours. For highly adherent crops, local watershed lines are formed through extreme erosion and morphological reconstruction operation control, which can better achieve the separation of adherent crops. However, because the false contours formed by local watershed lines cannot depict the real edges between the adhering crops well, the phenomenon of "straight edge" watersheds occurs. The Mask R-CNN model has less over- and under-segmentation and can restore the true edges of each cabbage under different degrees of adhesion.

Table 7 shows the counting accuracy of the OBIA-MDTWS and Mask R-CNN methods for cabbages planted in the entire study area. It can be observed that the images processed by the OBIA-MDTWS model have

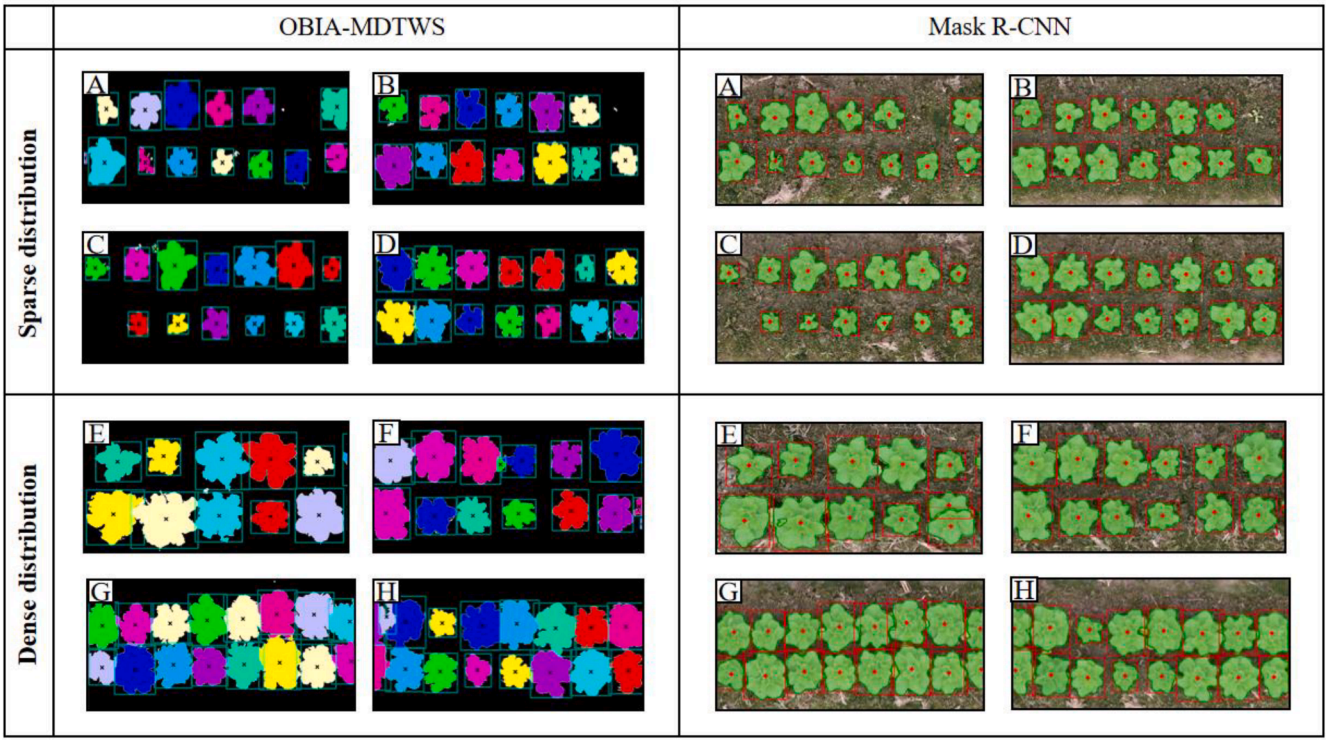


Fig. 13. Individual detection results of cabbage with different plant density by OBIA-MDTWS and Mask R-CNN methods.

Table 7
Accuracy of different methods.

Method	The ground truth	Predicted	True	False	Miss	False detection rate (%)	Miss detection rate (%)	Accuracy (%)
OBIA-MDTWS	1410	1375	1311	41	59	2.91	4.18	95.35
Mask R-CNN	1410	1429	1403	19	0	1.35	0	99.50

Table 8
Comparison of efficiency of different methods.

Method	Steps	Time (s)
OBIA-MDTWS	Image segmentation	12,488
	Train	15,269
	Classification	34,635
	Detection	209
	Total	62,601
Mask R-CNN	Train	16,561
	Prediction	178
	Total	16,739

more detection errors in both sparse-distribution and dense-distribution areas than the Mask R-CNN model. For the OBIA-MDTWS, the number of false detection was 41, the number of miss detection was 59, the false detection rate was 2.91 %, the miss detection rate was 4.18 %, and the overall accuracy rate was 95.35 %. In contrast, Mask R-CNN performed better overall, with only 19 false detection strains and no missed detection. Compared with the OBIA-MDTWS, false detection rate decreased by 1.56 percentage points, the missed detection rate decreased by 4.18 percentage points, and its overall accuracy reached 99.50. The results suggest that the deep learning model (Mask R-CNN) is superior to the OBIA-MDTWS method in segmentation and counting, and has higher robustness and generalization.

3.4. Comparison of efficiency of different methods

To explore the recognition efficiency of each method on the cabbage detection task, we used two methods, Mask R-CNN and OBIA-MDTWS, to conduct three experiments on the images in the study area. The results (Table 8) showed that the time consumption of the OBIA-MDTWS is significantly higher than that of the Mask R-CNN deep-learning model. The overall average running time was 62,601 s, of which image segmentation, ML model training, and classification times were the highest, with 12,488 s, 15,269 s, and 34,635 s, respectively. This indicates that the OBIA method has certain limitations in processing high-resolution images such as high computational cost and low efficiency. However, for the Mask R-CNN deep learning model, only the training session consumed a relatively high proportion of time, that is, 16,561 s. The inference and prediction results session consumed only 178 s, and had significant speed advantage over the OBIA-MDTWS method. The time is shortened by 45,862 s, and the speed is increased by 3.74 times. In summary, the Mask R-CNN deep learning model exhibits significant potential for real-time detection.

4. Discussion

In this study, we adopted two mainstream detection methods, deep learning (Mask R-CNN) and the OBIA-MDTWS algorithm, to extract and estimate the number of cabbages. Both methods performed well; however, there were certain differences in flexibility and functionality. We systematically analyzed and compared the two methods in terms of their process flow, recognition accuracy, robustness, and efficiency. The findings are as follows.

The OBIA-MDTWS method consisting of several independent modules, performs cabbage extraction and individual detection through the processes of “image segmentation → feature extraction → object

classification → target detection" (Nababan et al., 2021). Several factors would affect the classification accuracy in the process: segmentation parameters, feature selection, classification algorithms, and watershed algorithm parameters. First, there is no absolute optimal scale parameter for different types of targets constituting an image scene. Although ESP2 can systematically evaluate the segmentation scale, it must be visually judged to attain the absolute optimal scale. This means that optimal scaling selection based on ESP2 cannot completely avoid subjectivity. It is important to emphasize that GA also requires a manual definition of the minimum/maximum values for each image segmentation parameter for each algorithm, which is essential to achieve good final output, minimum running time, and appropriate parameter combinations. Therefore, extensive experimentation and experience are required to obtain better parameter values. Moreover, when the spectral, geometric, and textural features of different targets are similar, it is difficult to obtain an ideal segmentation result, which impacts the subsequent classification accuracy. Second, extracting and optimizing multiple feature variables of an image object is a crucial step in the OBIA-MDTWS method. Too many unnecessary features will lead to multicollinearity and redundancy among features (Sumesh et al., 2021), thus increasing the complexity of the training process, reducing the generalization ability and accuracy of the algorithm, leading to deviations in classification accuracy. Traditional classification requires manual design of operators for feature extraction (Wang et al., 2021). However, selecting suitable and comprehensive features is usually difficult, which limits the applicability of the OBIA-MDTWS algorithm. Therefore, certain feature-screening algorithms must be used to obtain the best combination of features. No particular combination of image segmentation and classification algorithms is suitable for all tasks/targets, implying that these algorithms "cannot be universally good" across all problems and domains, especially image segmentation algorithms. This means that comparing and optimizing procedures is crucial to determine which are better suited for a particular target. Our results showed that the multiresolution segmentation algorithm combined with the RF classifier exhibits better overall performance in various tests, which is consistent with previous findings (Fernandez-Delgado et al., 2014). However, this may result from the algorithm requiring less fine-tuning than other algorithms (e.g., SVM and KNN) (Goncalves et al., 2019). Finally, although the improved distance-transformed watershed algorithm proposed in this paper can effectively solve the overlapping adhesion problem during the segmentation of high-density cabbage individuals, the algorithm parameters must be adjusted according to the background noise, individual distribution density, and shape characteristics at the time of image acquisition to achieve the optimal extraction effect. Planting areas with different growth states and adhesion conditions cannot be segmented and detected using uniform parameters for single cabbage plants. Planting areas with different growth states and adhesion conditions cannot use unified parameters to segment and detect individual cabbages. Therefore, when the OBIA-MDTWS uses multiple modules to solve complex tasks, there is an apparent disadvantage that the training objectives of each module are inconsistent. It is impossible to ensure that the output results of each module reach the optimal value, resulting in final deviations from the macroscopic objectives of the system, which makes it difficult to achieve optimal performance. Another problem is that the accumulation of errors and the deviation generated by the former module may affect the latter, resulting in the inability to reduce the overall intrinsic error to a minimum.

Unlike the OBIA-MDTWS algorithm, the Mask R-CNN deep learning method is an end-to-end algorithm that controls the process of "feature extraction → classification" using the image information as a supervised signal (Hou et al., 2021). It combines the three stages and does not require the selection of segmentation parameters. Therefore, it can avoid the problems of under-segmentation and over-segmentation of the segmented objects. There is no need to extract image features manually, which can avoid the complexity and inefficiency of feature selection and classification inaccuracy caused by shallow classification algorithms (Zhang and Chi, 2020). It realizes an efficient and high-precision classification of remote sensing images in the form of end-to-end pixels. The "end-to-end" learning approach has the advantage of synergy and is more likely to obtain the global optimal solution. The classification results were significantly better than those obtained using the traditional object-oriented ML methods. This is consistent with the results of the improved FCN (Ghorbanzadeh et al., 2022), which had a higher accuracy rate for field crop recognition than the classical object-oriented SVM method. Because the existing deep learning model is complex with the internal decision-making method opaque, poorly interpretable or un-interpretable, it appears like a "black box". Although it is possible to find which nodes of deep neural networks are activated using mathematics, it is not possible to determine the neurons responsible for it. Hence, the relationship between the results and influencing factors cannot be explained further. Therefore, the deep learning model has poor transferability and its application is restricted. In contrast, machine learning algorithms such as RF, SVM, and KNN provide us with clear rules, indicating what to choose and why to choose; hence, it is easy to explain the reasoning process behind the algorithm. Therefore, despite the high accuracy of deep learning models, researchers still tend to select machine learning algorithms for remote sensing image information extraction tasks that must be interpretable.

Meanwhile, for the detection of an individual in a segmented image, the Mask R-CNN model is first trained based on a large number of labeled samples, and then the trained model is used to segment single individuals. It relies less on a priori knowledge and does not require the debugging of specific regions. It can automatically complete the learning process based on a given dataset and can realize the extraction of cabbage under different lighting conditions, growth states, and plant densities. It has a strong generalization ability and is a low-cost and highly-efficient vegetable extraction method based on UAV visible-light images. This method can provide real-time and effective data for field crop management in precision agriculture.

In addition, owing to the ultra-high spatial resolution of UAV images (pixel of 1.56 mm × 1.56 mm), the OBIA-MDTWS method needs to calculate and segment a large number of homogeneous image objects (a total of 346,884 object units were segmented in this study) in early stage. In the process of classification and application, it is necessary to use ML algorithms to analyze the multi-dimensional features of each object comprehensively. This process can use only the central processing unit (CPU) for operation. Although the CPU is fast in executing a single process, it requires multiple rounds of operations to calculate all object attributes, significantly increasing the cost of computing time. Therefore, all links consumed more time than the deep learning model method used in this study. In contrast, the deep learning model uses a pixel-end-to-pixel-end form of computation and graphics processor (GPU) for computing. Because GPUs provide the infrastructure for multi-core parallel computing and have several cores, they can support the parallel computation of large amount of data, thus solving large and complex neural network inference problems. Therefore, the recognition efficiency of deep learning is significantly higher than that of the OBIA-MDTWS method, and it can perform real-time extraction and quantity estimation.

However, because the sample dataset we selected was restricted to a specific time and space, the robustness of the two models in dealing with UAV images of different spectral, spatial, and temporal resolutions has not been tested. In addition, our study did not consider the impact of non-remote sensing data (such as meteorological data) or the different stages of the phenological period on the classification results. Therefore, future research will test the methods using multiscale, multi-sensor, multitemporal, and multiregional UAV images. Moreover, we will constantly optimize the model parameters to improve the structure, feature extraction, and data mining capabilities, and further compare the spatiotemporal generalization capabilities of the models.

5. Conclusion

In this study, we investigated the capabilities of pixel-based deep learning and object-based image analysis for individual detection of cabbage plants based on UAV images. Experimental results have verified the effectiveness of the Mask R-CNN deep learning model for cabbage extraction, with an overall average F1-score of 97.63 %, which is higher than that of the OBIA-MDTWS algorithm. Mask R-CNN can generate high-quality segmentation masks and bounding boxes for each target. The overall accuracy in estimating the number of cabbages in the study area was 99.50 %, which is 4.15 percentage points higher than that of the OBIA-MDTWS algorithm. Moreover, the Mask R-CNN method uses GPU acceleration and other means to optimize the recognition speed of the model, and its running time is significantly shorter than that of the OBIA-MDTWS algorithm that integrates multiple modules. In summary, the Mask R-CNN deep learning method can more accurately and effectively monitor crop growth information in artificially planted fields, and its integration with the UAV remote sensing platform have broader application prospects.

CRediT authorship contribution statement

Zhangxi Ye: Conceptualization, Methodology, Software, Resources, Writing – original draft, Writing – review & editing, Visualization. **Kaile Yang:** Conceptualization. **Yuwei Lin:** . **Shijie Guo:** Software, Validation, Formal analysis. **Yiming Sun:** Writing – original draft, Visualization. **Xunlong Chen:** Methodology, Investigation. **Riwen Lai:** . **Houxi Zhang:** Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was supported by the Tibet Autonomous Region Science and Technology Plan Project Key Project (XZ202201ZY0003G), Forestry Peak Discipline Construction Project of Fujian Agriculture and Forestry University (72202200205), National Natural Science Foundation of China (31901298), and the Natural Science Foundation of Fujian Province (2021 J01059).

References

Ahansal, Y., Bouziani, M., Yaagoubi, R., Sebari, I., Sebari, K., Kenny, L., 2022. Towards smart irrigation: a literature review on the use of geospatial technologies and machine learning in the management of water resources in arboriculture. Agronomy 12 (2), 297. https://doi.org/10.3390/agronomy12020297.

Bian, L.M., Zhang, H.C., Ge, Y.F., Cepl, J., Stejskal, J., El-Kassaby, Y.A., 2022. Closing the gap between phenotyping and genotyping: review of advanced, image-based phenotyping technologies in forestry. Ann. For. Sci. 79 (1), 21. https://doi.org/10.1186/s13595-022-01143-x.

Bolya, D., Zhou, C., Xiao, F.Y., Lee, Y.J., & Ieee 2019. YOLACT real-time instance segmentation. In, IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9156-9165). Seoul, SOUTH KOREA. https://doi.org/10.1109/iccv.2019.00925.

Chen, L.C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H., & Ieee 2018. MaskLab: instance segmentation by refining object detection with semantic and direction features. In, 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4013-4022). Salt Lake City, UT. 10.48550/arXiv.1712.04837.

Cobelo, I., Machado, K.B., David, A.C.M., Carvalho, P., Ferreira, M.E., Nabout, J.C., Unmanned aerial vehicles and low-cost sensor as tools for monitoring freshwater chlorophyll-a in mesocosms with different trophic state. Int. J. Environ. Sci. Technol., 12. 10.1007/s13762-022-04386-3.

Colpaert, A., 2022. Satellite and UAV platforms, remote sensing for geographic information systems. Sensors 22 (12), 2. https://doi.org/10.3390/s22124564.

Csillik, O., Cherbini, J., Johnson, R., Lyons, A., Kelly, M., 2018. Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks. Drones 2 (4), 16. https://doi.org/10.3390/drones2040039.

Evagorou, E., Argyriou, A., Papadopoulos, N., Mettas, C., Alexandrakis, G., Hadjimitsis, D., 2022. Evaluation of satellite-derived bathymetry from high and medium-resolution sensors using empirical methods. Remote Sens. 14 (3), 19. https://doi.org/10.3390/rs14030772.

Fan, D., Su, X., Weng, B., Wang, T., Yang, F., 2021. Research progress on remote sensing classification methods for farmland vegetation. AgriEngineering 3 (4), 971–989. https://doi.org/10.3390/agriengineering3040061.

Feizi, A., Vahabzadeh, Z., Maleki, V., 2022. Evaluation of drought impacts on vegetation changes trend using hydrological re-analysis models and remote sensing. Water Resour. 49 (4), 689–698. https://doi.org/10.1134/s0097807822040054.

Feng, H.K., Tao, H.L., Li, Z.H., Yang, G.J., Zhao, C.J., 2022. Comparison of UAV RGB imagery and hyperspectral remote-sensing data for monitoring winter wheat growth. Remote Sens. 14 (15), 22. https://doi.org/10.3390/rs14153811.

Feng, Q., Yang, J., Liu, Y., Ou, C., Zhu, D., Niu, B., Liu, J., Li, B., 2020. Multi-temporal unmanned aerial vehicle remote sensing for vegetable mapping using an attention-based recurrent convolutional neural network. Remote Sens. 12 (10) https://doi.org/10.3390/rs12101668.

Fernandez-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach Learn Res. 15, 3133–3181. https://doi.org/10.5555/2627435.2697065.

Filippi, A.M., Guneralp, I., Castillo, C.R., Ma, A.D., Paulus, G., Anders, K.H., 2022. Comparison of image endmember- and object-based classification of very-high-spatial-resolution unmanned aircraft system (UAS) narrow-band images for mapping riparian forests and other land covers. Land 11 (2), 33. https://doi.org/10.3390/land11020246.

Gao, H., He, L., He, Z.W., Bai, W.Q., 2022. Early landslide mapping with slope units division and multi-scale object-based image analysis - a case study in the Xianshui river basin of Sichuan. China. J. Mt. Sci. 19 (6), 1618–1632. https://doi.org/10.1007/s11629-022-7333-6.

Ghorbanzadeh, O., Shahabi, H., Crivellari, A., Homayouni, S., Blaschke, T., Ghamisi, P., 2022. Landslide detection using deep learning and object-based image analysis. Landslides 19 (4), 929–939. https://doi.org/10.1007/s10346-021-01843-x.

Goncalves, J., Pocas, I., Marcos, B., Mucher, C.A., Honrado, J.P., 2019. SegOptim-A new package for optimizing object-based image analyses of high-spatial resolution remotely-sensed data. Int. J. Appl. Earth Obs. Geoinf. 76, 218–230. https://doi.org/10.1016/j.jag.2018.11.011.

Guo, Q., Zhang, J., Guo, S., Ye, Z., Deng, H., Hou, X., Zhang, H., 2022. Urban tree classification based on object-oriented approach and random forest algorithm using unmanned aerial vehicle (UAV) multispectral imagery. Remote Sens. 14 (16), 3885. https://doi.org/10.3390/rs14163885.

Gurunathan, A., Krishnan, B., 2022. A hybrid CNN-GLCM classifier for detection and grade classification of brain tumor. Brain Imaging Behav. 16 (3), 1410–1427. https://doi.org/10.1007/s11682-021-00598-2.

Han, Y.F., Wang, P., Zheng, Y.G., Yasir, M., Xu, C.M., Nazir, S., Hossain, M.S., Ullah, S., Khan, S., 2022b. Extraction of landslide information based on object-oriented approach and cause analysis in Shuicheng. China. Remote Sens. 14 (3), 30. https://doi.org/10.3390/rs14030502.

Han, Q.Z., Yin, Q., Zheng, X., Chen, Z.Y., 2022a. Remote sensing image building detection method based on mask R-CNN. Complex Intell. Syst. 8 (3), 1847–1855. https://doi.org/10.1007/s40747-021-00322-z.

He, K., Gkioxari, G., Dollar, P., Girshick, R., & Ieee 2017. Mask R-CNN. In, 16th IEEE International conference on computer vision (ICCV) (pp. 2980-2988). Venice, ITALY. 10.48550/arXiv.1703.06870.

Hou, F.F., Lei, W.T., Li, S., Xi, J.C., Xu, M.D., Luo, J.B., 2021. Improved mask R-CNN with distance guided intersection over union for gpr signature detection and segmentation. Autom. Constr. 121, 14. https://doi.org/10.1016/j.autcon.2020.103414.

Jiang, Y., Li, C., Paterson, A.H., Robertson, J.S., 2019. DeepSeedling: deep convolutional network and kalman filter for plant seedling detection and counting in the field. Plant Methods. 15, 141. https://doi.org/10.1186/s13007-019-0528-3.

Jiang, L.G., Wu, S., Liu, Y., 2022. Change analysis on the spatio-temporal patterns of main crop planting in the middle Yangtze Plain. Remote Sens. 14 (5), 26. https://doi.org/10.3390/rs14051141.

Kamarulzaman, A.M.M., Jaafar, W., Maulud, K.N.A., Saad, S.N.M., Omar, H., Mohan, M., 2022. Integrated segmentation approach with machine learning classifier in detecting and mapping post selective logging impacts using UAV imagery. Forests 13 (1), 24. https://doi.org/10.3390/f13010048.

Kamga, G.A.F., Bitjoka, L., Akram, T., Mbom, A.M., Naqvi, S.R., Bouroubi, Y., 2021. Advancements in satellite image classification : methodologies, techniques, approaches and applications. Int. J. Remote Sens. 42 (20), 7662–7722. https://doi.org/10.1080/01431161.2021.1954261.

Komarek, J., Klapst, P., Hrach, K., Kloucek, T., 2022. The potential of widespread UAV cameras in the identification of conifers and the delineation of their crowns. Forests 13 (5), 14. https://doi.org/10.3390/f13050710.

Kumari, K.S., Haleem, S.L.A., Shivaprakash, G., Saravanan, M., Arunsundar, B., Pandraju, T.K.S., 2022. Agriculture monitoring system based on internet of things by deep learning feature fusion with classification. Comput. Electr. Eng. 102, 14. https://doi.org/10.1016/j.compeleceng.2022.108197.

Li, Y.Q., Bai, J.W., Zhang, L., Yang, Z.H., 2022. Mapping and spatial variation of seagrasses in Xincun, Hainan Province, China. Based on Satellite Images. Remote Sens. 14 (10), 24. https://doi.org/10.3390/rs14102373.

Li, L.H., Jing, W.P., Wang, H.H., 2021. Extracting the forest type From remote sensing images by random forest. IEEE Sens. J. 21 (16), 17447–17454. https://doi.org/10.1109/jsen.2020.3045501.

Liu, P.F., Wang, Q., Yang, G.C., Li, L., Zhang, H., 2022. Survey of road extraction methods in remote sensing images based on deep learning. PFG-J. Photogramm. Remote Sens. Geoinf. Sci. 90 (2), 135–159. https://doi.org/10.1007/s41064-022-00194-z.

Ma, Z.G., Guo, Q.Y., Yang, F.Y., Chen, H.L., Li, W.Q., Lin, L.L., Zheng, C.Y., 2021. Recent changes in temperature and precipitation of the summer and autumn seasons over Fujian Province. China. Water 13 (14), 15. https://doi.org/10.3390/w13141900.

Mavridou, E., Vrochidou, E., Papakostas, G.A., Pachidis, T., Kaburlasos, V.G., 2019. Machine vision systems in precision agriculture for crop farming. J Imaging 5 (12). https://doi.org/10.3390/jimaging5120089.

Mehmood, K., Bao, Y.S., Mushtaq, S., Saifullah Khan, M.A., Siddique, N., Bilal, M., Heng, Z., Huan, L.i., Tariq, M., Ahmad, S., 2022. Perspectives from remote sensing to investigate the COVID-19 pandemic: a future-oriented approach. Front. Public Health 10. https://doi.org/10.3389/fpubh.2022.938811.

Nababan, B., Mastu, L.K., Idris, N.H., Panjaitan, J.P., 2021. Shallow-Water Benthic Habitat Mapping Using Drone with Object Based Image Analyses. Remote Sensing 13 (21), 23. https://doi.org/10.3390/rs13214452.

Okorie, A., Makrogiannis, S., 2019. Region-based image registration for remote sensing imagery. Comput. Vis. Image Underst. 189, 15. https://doi.org/10.1016/j.cviu.2019.102825.

Padua, L., Matese, A., Di Gennaro, S.F., Morais, R., Peres, E., Sousa, J.J., 2022. Vineyard classification using OBIA on UAV-based RGB and multispectral data: a case study in different wine regions. Comput. Electron. Agric. 196, 15. https://doi.org/10.1016/j.compag.2022.106905.

Pantic, I., Dacic, S., Brkic, P., Lavrnja, I., Jovanovic, T., Pantic, S., Pekovic, S., 2015. Discriminatory ability of fractal and grey level co-occurrence matrix methods in structural analysis of hippocampus layers. J. Theor. Biol. 370, 151–156. https://doi.org/10.1016/j.jtbi.2015.01.035.

Paoletti, M.E., Haut, J.M., Alipour-Fard, T., Roy, S.K., Hendrix, E.M.T., Plaza, A., 2022. Separable attention network in single- and mixed-precision floating point for land-cover classification of remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 5. https://doi.org/10.1109/lgrs.2021.3108965.

Pendyala, G., Kalluri, H.K., Rao, V.C., 2021. An efficient multi-stage object-based classification to extract urban building footprints from HR satellite images. Trait. Signal 38 (1), 191–196. https://doi.org/10.18280/ts.380120.

Qin, R.J., Liu, T., 2022. A review of landcover classification with very-high resolution remotely sensed optical images-analysis unit, model scalability and transferability. Remote Sens. 14 (3), 28. https://doi.org/10.3390/rs14030646.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Shi, P.F., Jiang, Q.G., Shi, C., Xi, J., Tao, G.F., Zhang, S., Zhang, Z.C., Liu, B., Gao, X., Wu, Q., 2021. Oil well detection via large-scale and high-resolution remote sensing images based on improved YOLO v4. Remote Sens. 13 (16), 19. https://doi.org/10.3390/rs13163243.

Siljeg, A., Panda, L., Domazetovic, F., Maric, I., Gasparovic, M., Borisov, M., Milosevic, R., 2022. Comparative assessment of pixel and object-based approaches for mapping of olive tree crowns based on UAV multispectral imagery. Remote Sens. 14 (3), 18. https://doi.org/10.3390/rs14030757.

Sumesh, K.C., Ninsawat, S., Som-ard, J., 2021. Integration of RGB-based Vegetation Index, Crop Surface Model and Object-based Image Analysis Approach for Sugarcane Yield Sstimation Using Unmanned Aerial Vehicle. Comput. Electron. Agric. 180, 19. https://doi.org/10.1016/j.compag.2020.105903.

Wang, M., Cui, Q., Sun, Y.J., Wang, Q., 2018. Photovoltaic panel extraction from very high-resolution aerial imagery using region-line primitive association analysis and template matching. ISPRS J. Photogramm. Remote Sens. 141, 100–111. https://doi.org/10.1016/j.isprsjprs.2018.04.010.

Wang, X., Zhang, R., Kong, T., Li, L., & Shen, C.J.A.i.N.i.p.s. 2020. Solov2: Dynamic and fast instance segmentation, 33, 17721-17732. 10.48550/arXiv.2003.10152.

Wang, Y., Qin, Y., Cui, J., 2021. Occlusion robust wheat ear counting algorithm based on deep learning. Front Plant Sci. 12, 645899 https://doi.org/10.3389/fpls.2021.645899.

Wu, Q.F., Feng, D.Q., Cao, C.Q., Zeng, X.D., Feng, Z.J., Wu, J., Huang, Z.Q., 2021. Improved mask R-CNN for aircraft detection in remote sensing images. Sensors 21 (8), 13. https://doi.org/10.3390/s21082618.

Xu, C.J., Zhu, G.B., Shu, J.Q., 2022. A combination of lie group machine learning and deep learning for remote sensing scene classification using multi-layer heterogeneous feature extraction and fusion. Remote Sens. 14 (6), 26. https://doi.org/10.3390/rs14061445.

Xue, Y.A., Zhao, J.L., Zhang, M.M., 2021. A watershed-segmentation-based improved algorithm for extracting cultivated land boundaries. Remote Sens. 13 (5), 19. https://doi.org/10.3390/rs13050939.

Yang, Z.Y., Yu, X.Y., Dedman, S., Rosso, M., Zhu, J.M., Yang, J.Q., Xia, Y.X., Tian, Y.C., Zhang, G.P., Wang, J.Z., 2022b. UAV remote sensing applications in marine monitoring: knowledge visualization and review. Sci. Total Environ. 838, 23. https://doi.org/10.1016/j.scitotenv.2022.155939.

Yang, K.L., Zhang, H.X., Wang, F., Lai, R.W., 2022a. Extraction of broad-leaved tree crown based on UAV visible images and OBIA-RF model: a case study for chinese olive trees. Remote Sens. 14 (10), 23. https://doi.org/10.3390/rs14102469.

Ye, Z.X., Wei, J.H., Lin, Y.W., Guo, Q., Zhang, J., Zhang, H.X., Deng, H., Yang, K.J., 2022. Extraction of olive crown based on UAV visible images and the U-2-Net deep learning model. Remote Sens. 14 (6), 20. https://doi.org/10.3390/rs14061523.

Zhang, Y.F., Chi, M.M., 2020. Mask-R-FCN: a deep fusion network for semantic segmentation. IEEE Access 8, 155753–155765. https://doi.org/10.1109/access.2020.3012701.

Zhang, S., Dai, X., Li, J., Gao, X., Zhang, F., Gong, F., Lu, H., Wang, M., Ji, F., Wang, Z., Peng, P., 2022a. Crop classification for UAV visible imagery using deep semantic segmentation methods. Geocarto Int. 37 (25), 10033–10057.

Zhang, W.B., Gao, F., Jiang, N., Zhang, C., Zhang, Y.C., 2022b. High-temporal-resolution forest growth monitoring based on segmented 3D canopy surface from UAV aerial photogrammetry. Drones 6 (7), 18. https://doi.org/10.3390/drones6070158.