International Journal of Applied Earth Observations and Geoinformation 106 (2022) 102662

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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

Counting trees in a subtropical mega city using the instance segmentation method

Ying Sun [a], Ziming Li [a], Huagui He [b], Liang Guo [b], Xinchang Zhang [c,*], Qinchuan Xin [a,*]

[a] School of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-Sen University, Guangzhou 510275, China
[b] Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 510060, China
[c] School of Geographical Science, Guangzhou University, Guangzhou 510006, China

ARTICLE INFO

Keywords:
Counting trees
Cascade mask R-CNN
Very high-spatial-resolution aerial images
Individual tree detection
Tree density
Guangzhou

ABSTRACT

Counting the trees in a given area or city is meaningful when making decisions for government policies and administration, including the international afforestation effort (i.e., the Trillion Trees Campaign). Determining individual trees on a large scale poses significant challenges, especially in subtropical and tropical areas, because of their diverse crown characteristics. Using very high-spatial-resolution images, we can see the tree crowns clearly. In this study, we counted the population number of trees in the subtropical mega city of Guangzhou, and we used an end-to-end tree-counting deep-learning framework in the regional-scale tree detection by delineating each tree crown. It is a simple framework in which individual trees can be detected directly without manual operation. We used the cascade mask regions with convolutional neural networks (CMask R-CNN) as the backbone and added three types of attention modules to build the derivatives of the CMask R-CNN. The experimental results showed that the CMask R-CNN performed the best among all of the methods, and more than 112 million individual trees with crown sizes of large than 1 m^2 were detected. The experimental assessment indicated that the accuracy was 88.32% in terms of the R^2 value and 82.56% in terms of the F1-score. This study not only revealed the number of trees, but it also provided the tree density at different scales, which is a prominent component of the ecosystem structure. We also analyzed the tree density at the 30 m and 1000 m scales. The experimental results showed that Guangzhou has a high canopy cover with a tree density of 150 trees per hectare. For the entire city, the tree density is highest in the northern area of Guangzhou, followed by the central part. From the central-western part to the central-eastern part, the tree density increases. The lowest tree density is located in the southern part of Guangzhou near the Pearl River estuary, which is a basic farmland conservation area. It is notable that the tree density of the urban land is about 15 trees per 900 m^2, indicating a good living environment. The method developed in this study provides a flexible means of large-scale tree counting without manual operation based on very high-spatial-resolution images.

1. Introduction

Trees contribute extensively to the biogeochemical cycles (Crowther et al., 2014; Hansen et al., 2013), and they are considered to be the best carbon capture and sequestration carriers in history, so they play an important role in reducing greenhouse gas emissions and mitigating the risk of climate change (Crowther et al., 2015). Currently, most previous research has focused on trees in forest area (Brandt et al., 2020; Pan et al., 2011). However, the urban carbon cycle is a popular issue in the fields of international politics and economics. Each tree is an important element of the vegetation ecosystem (Duinker et al., 2015; Romero-Lankao et al., 2018). Ascertaining the number of trees in a given area or city is meaningful when making decisions about government policies and administration. This number can provide the basis for forest inventory and carbon sequestration capacity assessments and holds great significance for incorporating urban trees into regional climate action plans and the carbon offset market as well as for promoting sustainable urban development. In addition, it is important for international afforestation efforts, such as the Trillion Trees Campaign launched at the World Economic Forum in 2020. Such projects need a baseline of the current number of trees in a certain area or city to establish the target and evaluate the result of the task (Crowther et al., 2015).

* Corresponding authors.
E-mail addresses: zhangxc@gzhu.edu.cn (X. Zhang), xinqinchuan@mail.sysu.edu.cn (Q. Xin).

https://doi.org/10.1016/j.jag.2021.102662
Received 10 July 2021; Received in revised form 12 December 2021; Accepted 20 December 2021
Available online 22 December 2021

Remote sensing makes large-scale Earth observation available, and the large number of forest cover studies have produced many data products (Pfeifer et al., 2012; Tuanmu and Jetz, 2014). However, the problem of counting the tree population is also relevant to many applications, such as forest inventories and environmental protection. With the increasing availability of high- and very high-spatial-resolution remote sensing data, we can gather information at the individual tree level (Qiu et al., 2020). Based on the spectral and textural features of high-resolution optical images or the elevation features of the point cloud, numerous automatic methods, such as watershed segmentation (Chen et al., 2006), region growing (Erikson, 2003), polynomial fitting (Wu et al., 2019), distance discriminant clustering (Li et al., 2012), adaptive mean shift (Yan et al., 2020), template matching (Vibha et al., 2009), and object-oriented image segmentation (Qiu et al., 2020) have been developed to detect individual trees in temperate forests (Duncanson et al., 2014). Rizeei et al. (2018) combined object-oriented image segmentation and regression analysis to count oil palms based on WorldView-3 and Light Detection and Ranging (LiDAR) data. Norzaki and Tahar (2019) reported that template matching performed better than other segmentation approaches. These studies often were based on small datasets and were conducted in small areas (Stereńczak et al., 2020; Zhou et al., 2017). Recently, Brandt et al. (2020) attempted to map the individual trees in the West African Sahara, which has a large area of 1.3 million km^2. Their work involved processing more than 50,000 VHR satellite images (i.e., a very large dataset), requiring a powerful computing ability. Although extensive efforts have been made, individual tree detection for forest inventories on a large scale remain a major challenge, especially in the subtropical and tropical areas. In these high-density subtropical and tropical areas, trees may have diverse architecture and crown characteristics (Wagner et al., 2018), such as crowns covered by lianas and new leaves on the crowns of evergreen trees, making individual tree detection much more difficult. This task, however, is not impossible, merely challenging. Khan and Gupta (2018) compared different tree-counting methods and found that deep learning is a quick and effective method.

Deep-learning networks have been popular since the deep belief network was first proposed in 2006 (Hinton et al., 2006). Over the past decade, deep learning has been demonstrated to have a mature and reliable ability in image classification (Chan et al., 2015), semantic segmentation (Long et al., 2015), and object detection (Lin et al., 2017) with the development of the Graphics Processing Unit (GPU), as well as public datasets, such as ImageNet. Image classification aims to assign a label to an image from a given dataset. Typical convolutional neural networks (CNNs) include the AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and DenseNet (Zhu and Newsam, 2017). With the development of deep networks, the number of CNN layers has increased from several to hundreds. Many architectures, such as the skip layer, dense connect, inception module, gate residual unit, and attention module, have been proposed to improve the feature extraction ability (Chen et al., 2019b). Zheng et al. (2020) developed a Multi-level Attention Domain Adaptation Network (MADAN) for counting oil palm trees based on high-resolution images. They used a classification method with lots of post-processing. Semantic segmentation is used to segment pixel regions containing different categories of objects and to determine their categories. The Fully Convolutional Network (FCN) (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), Unet (Ronneberger et al., 2015), Deeplab (Chen et al., 2017), and PSPNet (Chen et al., 2017) are commonly used networks. In addition, many derivatives of these FCNs have been developed for lightweight modeling (Chen et al., 2021a) and model performance improvement (Chen et al., 2021b). In this type of deep network, various types of up-sampling structures have been used to restore the compressed features from the CNN encoding to the same size as the input image. Yao et al. (2021) used four networks, including CNNs and FCNs, for tree counting using GF-II images. They reported that the encoder-decoder FCNs had a better accuracy than the CNNs. A similar study was conducted by Tong et al. (2021), and they counted trees using a point-wise supervised segmentation network. However, semantic segmentation cannot separate one object from another object in the same category. Object detection or recognition detects each object in the input image and assigns the corresponding object category. The object can be detected individually using a bounding box. The popular algorithms include Region-based Convolutional Neural Networks (RCNNs) (Girshick et al., 2014), SPPNet (He et al., 2015), and Fast RCNNs (Ren et al., 2015). These methods are two-step algorithms. First, they obtain the proposed regions, and then, they conduct bounding box refinement and category prediction. Yolo started the era of one-step object detection. Weinstein et al. (2020) proposed a DeepForest network for individual tree detection based on high-resolution red–greenblue (RGB) images, which is encouraging. Ammar et al. (2021) proposed a deep-learning framework for counting palm trees based on aerial geotagged images, in which three object detection networks were involved. However, the tree crowns could not be delineated. Instance segmentation combines the three types of tasks and can delineate the fine outline of each instance and the category (He et al., 2017). Counting trees through crown delineation is an instance segmentation task. Ocer et al. (2020) tried to detect the trees using the Mask R-CNN and a feature pyramid network (FPN). Their study area was a campus about 93 ha, in which most of the trees were isolated trees. Overall, most of these studies have tested their methods with a relatively open forest, such as oil palm trees, olive trees, and fruit trees (Osco et al., 2020; Salamí et al., 2019; Santoro et al., 2013), using high-resolution aerial/satellite images or LiDAR data.

Tropical and subtropical trees are an important component of global trees, accounting for about 42% of all trees worldwide (Crowther et al., 2015). The goal of this study was to count the trees in a subtropical mega city by delineating the crowns in very high-resolution (VHR) images. To achieve this goal, we developed a wall-to-wall tree delineation method using the cascade Mask R-CNN network (Chen et al., 2019a). Compared with the Mask R-CNN, it can perform cascaded refinement of the instance segmentation through joint multistage processing. The VHR images were cropped into small patches and directly fed into the cascade Mask R-CNN for the individual tree crown delineation. The performance was assessed using both the error matrix and the coefficient of determination. Furthermore, we analyzed the tree density at different scales and the characteristics of the tree density in different land use types based on the trees detected.

2. Study materials

2.1. Study area

In this study, we selected Guangzhou (centered at 23°06′32″N, 113°15′53″E) as the study area, which is the third largest city in China and is (Fig. 1). It contains 11 administrative districts, with an area of 7434.4 km^2, and is located on the subtropical coast and close to the South China Sea. It has a marine subtropical monsoon climate, which is characterized by a warm and rainy climate with an annual average temperature of 20–22 °C. Guangzhou is a hilly area with high terrain in the northeast and low terrain in the southwest. In the north, there is a forest concentrated hilly area, and the highest peak has an altitude of 1210 m. Abundant rainfall is conducive to the growth of plants, and the vegetation is green year-round. The trees in Guangzhou transition from the subtropical to tropical zone. Although the urbanization rate was 86.46% in 2020, 84 forest reserves or parks in Guangzhou were located in almost all of the districts, the largest of which is located in the district of Conghua.

2.2. Airborne optical images

The RGB images of Guangzhou City (22°26′–23°56′N, 112°57′–114°3′E) were acquired using a digital camera mounted on a

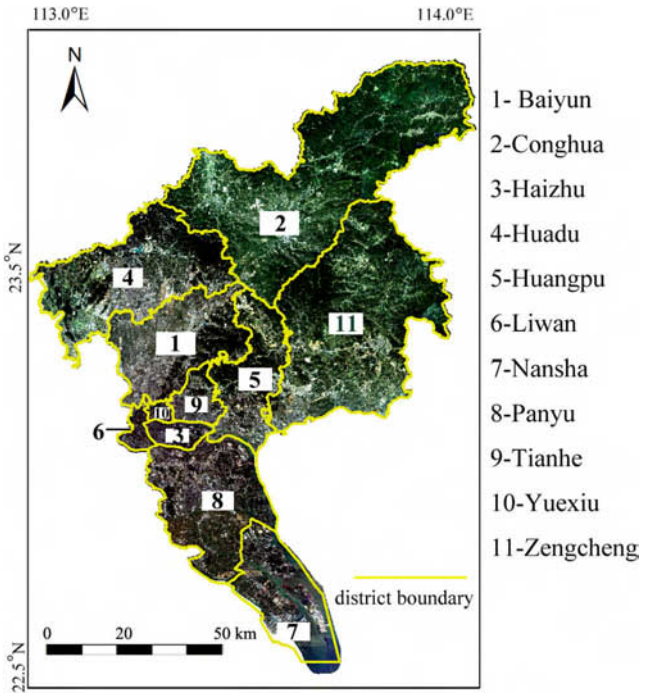

Fig. 1. The study area in Guangzhou, China.

Yun5 turboprop airplane under clear sky conditions. Because this is a large area, the data collection took a long time (from September 2017 to January 2019). The images were acquired in three commonly used spectral bands (i.e., RGB), with a spatial resolution of approximately 0.1 m. Uniform color operation was carried out when there was a color difference caused by the light conditions during the imaging. After this, the images were mosaicked into a complete image using the automatic seamless mosaic method. All of the images were orthorectified to the CGCS 2000 coordinate system. For the convenience of storage and operation, the image of Guangzhou was cropped into image tiles. A total of 7902 10,000 × 10,000 pixel image titles were obtained. We used all of the images in this study regardless of where the area was located (in the water or no tree cover).

2.3. Land cover maps of Guangzhou

For further analysis, we used the GlobeLand30 land cover products with a spatial resolution of 30 m for the land cover extraction of Guangzhou. GlobeLand30 was produced by the National Basic Geographic Information Center of China based on Landsat TM5, ETM+, and OLI and HJ-1 and GF-1 satellite images (http://www.globallandcover.com/). The products for both 2010 and 2020 were employed, and the land cover types were merged into six classes: farmland, forestland, grassland, water bodies, urban land, and other land. To investigate the change in the tree population, we derived a land cover change map of Guangzhou from the bitemporal land cover maps. Finally, the land cover maps were reprojected onto the CGCS 2000 coordinate system to match the airborne optical images and the tree detection results.

3. Methods

In this study, we conducted tree counting in the subtropical mega city of Guangzhou. We used 7902 airborne optical image titles, which each contained 10,000 × 10,000 pixels (1000 m × 1000 m), to ascertain the number of trees. First, we cropped the image titles into small patches according to the relative position. Second, the image patches were input into a modern instance segmentation network (Cascade Mask R-CNN) for training and testing to detect the individual tree crowns. The area of the tree canopy and the location (row and column) of each tree center were recorded. As the detected tree location was in the image patch without coordinate information, the detected trees were reprojected onto the original CGCS 2000 coordinate system for further assessment and analysis. The entire workflow was simple and easy to implement, which is a clear benefit for large-scale applications.

3.1. Sample labeling

Based on the GPU memory of our computer, we chose two patch sizes, 1000 × 1000 pixels and 1024 × 1024 pixels, as the model input sizes. We accounted for the different sceneries in the city in the tree sample selection, including urban land, parks, farmland, and forests, for the sample selection (Fig. 2), and the image patches were randomly selected from these scenes. As the trees are green all year long in Guangzhou, we did not consider the seasonal effect in the sample selection. A tree instance dataset was produced based on these very high-spatial-resolution aerial images and will be made public later. The training dataset contained 477 1024 × 1024 pixels image patches, and the test dataset contained 51 image patches of both sizes (1000 × 1000 pixels and 1024 × 1024 pixels). We labeled each tree crown manually for the instance segmentation using ArcGIS 10.4.1. We delineated the tree crowns based on visual interpretation. As the light cannot penetrate the canopy, we considered only the upper forest trees. Finally, 118,948 and 9021 tree crown instances were labeled for the model training and testing, respectively. To detect all of the trees, the 7902 airborne optical image tiles were cropped into smaller 1000 × 1000 pixels (100 m × 100 m) image patches, and 790,200 images were generated covering all of Guangzhou.

3.2. Tree counting based on Cascade Mask R-CNN

Instance segmentation has been improved significantly because of the public COCO dataset. The Cascade Mask R-CNN has been demonstrated to have an excellent ability in leveraging the reciprocal relationship between detection and segmentation (Chen et al., 2019a). Compared with the Mask R-CNN, it improves the bounding box regression and mask detection by the cascade and multi-task learning. As is shown in Fig. 3, we developed the tree crown delineation framework based on the Cascade Mask R-CNN. This framework allowed the input of an image patch of any size within the memory of the GPU, and we slightly modified the data input of the network to fit our dataset. Attention mechanisms have been shown to be helpful in information selection and feature representation in channel, spatial, and temporal wise based on the fusion of different level features. Here, we added three types of attention modules—that is, squeeze-and-excitation (SE) (Hu et al., 2018), convolutional block attention module (CBAM) (Woo et al., 2018), and Coordinate Attention block (CA) (Hou et al., 2021)—to the feature extraction part to achieve better tree crown detection. We also amended the output layer for the convenience of obtaining the statistics of the results. The proposed architecture begins with feature extraction, and we introduced ResNet50 and FPN (Lin et al., 2017) as the backbone network. The attention modules were inserted in the bottleneck of ResNet50 (Fig. 3). The feature extraction part reduced the dimension and extracted the useful features. With the attention module, different levels of spatial and semantic features could be obtained for the subsequent tasks. The Regional Proposal Network (RPN) generated the individual tree proposal regions (anchor boxes) based on the detected features from the backbone network. Moreover, the RPN distinguished the individual trees from the background (other objects) and refined the anchor boxes using the Intersection Over Union (IOU) between the anchor boxes and the ground truth. The refined anchor boxes from the RPN were combined with the feature maps (Regions Of Interest, ROIs) and then were input into the following procedures: individual tree detection (boundary box in Fig. 3) and tree crown segmentation (mask in Fig. 3).

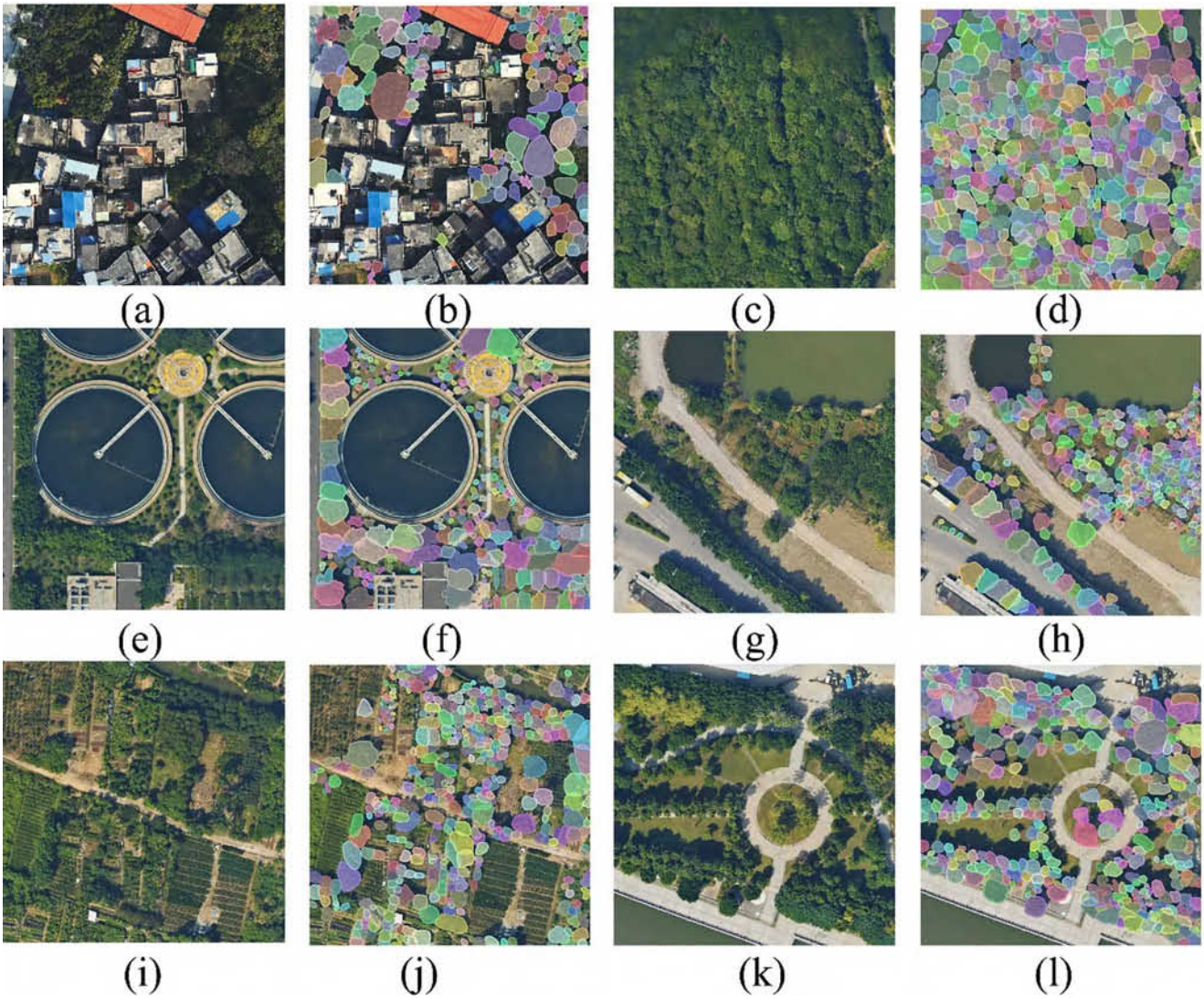

Fig. 2. Tree sample selection from different sceneries in the city, including (a) urban land, (c) forests, (e) factories, (g) lakeside, (i) farmland, and (k) parks. (b), (d), (f), (h), (j), and (l) are the corresponding labels for sceneries (a), (c), (e), (g), (i), and (k), respectively. The masks with different colors in the labeling are the individual tree crowns.

The feature maps were often smaller than the input images in terms of size. The ROI Align tool maps the ROIs to the location of the input image and makes them the same size. The position of the individual tree bounding box (green rectangles in Output in Fig. 3), the crown size, and the score of the detected instance were recorded. As the spatial reference was missed in the instance segmentation, we reprojected the detected instances onto the original coordinate system according to the relative location relationship. Compared with segmentation, the tree crowns could be delineated without post-processing.

The experiment was implemented in Ubuntu18.04. We used the Cascade Mask R-CNN in MMdetection 2.7 for the tree crown detection and built the attention derivatives, which were run on an NVIDIA GeForce 2080Ti with an 11 GB memory based on Pytorch1.4. The details of the network are shown in Table 1. The backbone of the ResNet50 and the neck of the FPN were used for the feature extraction, and the cascade ROI heads were then used for the tree location detection and tree crown delineation. The parameters were set as follows: the base learning rate was set as 0.02. The learning rate was adaptively updated using the warmup strategy for the first 1000 iterations, and the warmup ratio was 0.0001. After this, the learning rate was adjusted to 10% of the existing learning rate at 50, 80, and 100 epochs using the learning rate decay. There were a total of 110 epochs in the training phase. The momentum was set as 0.9, and the weight decay was 0.0001.

3.3. Assessment

To assess the performance of the attention guided Cascade Mask R-CNN, we also used the famous object detection model YOLO to compare the proposed method with other methods. Finally, five models (i.e., the Cascade Mask R-CNN, SE Cascade Mask R-CNN, CBAM Cascade Mask R-CNN, CA Cascade Mask R-CNN, and YOLO) were used to conduct tree counting in Guangzhou. We employed two types of assessments. One was the error matrix, which assesses the percentage of trees detected. For the 51 test images, which each corresponded to about a 100 m $\times$ 100 m plot, we counted the trees that were correctly detected (true positive, TP), incorrectly detected (false positive, FP), and undetected (false negative, FN); then, we calculated the precision, recall, F1 score, and detection rate (DR) (Yin and Wang, 2016). The closer the value of F1 was to 1, the better the counting effect was. The DR could be greater than 1 (overestimation), less than 1 (underestimation), and equal to 1 (good estimation).

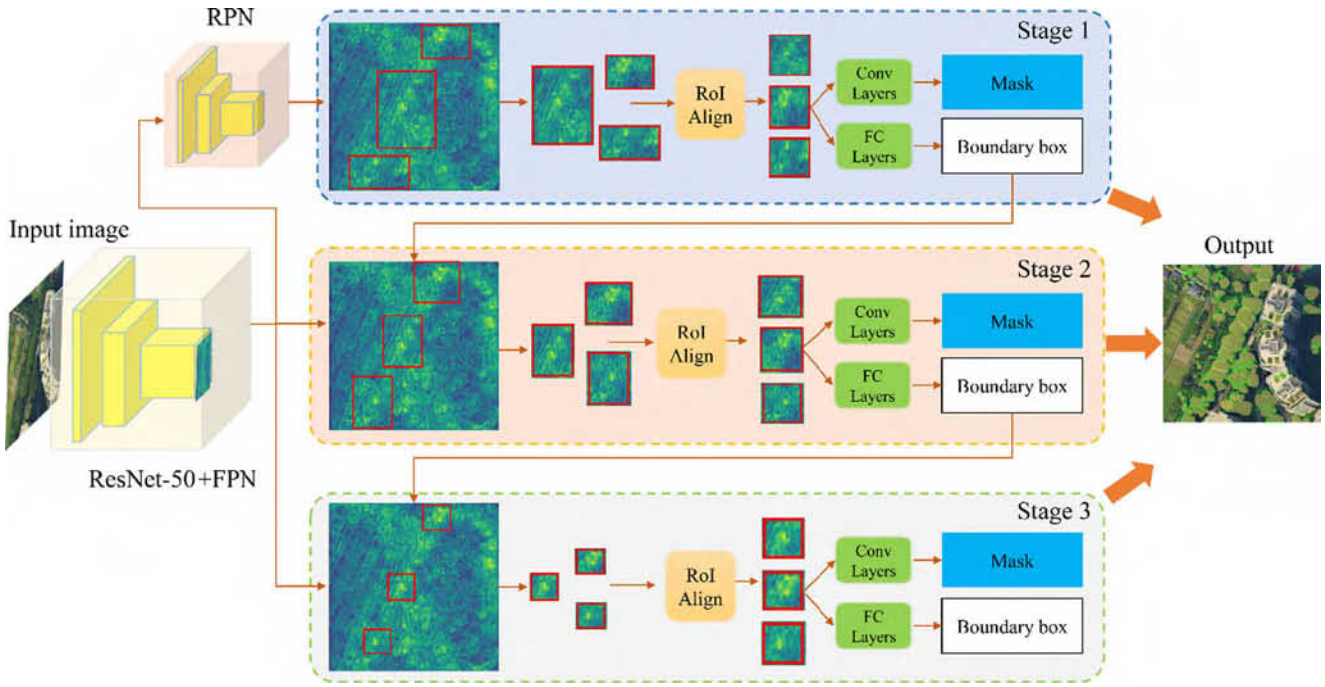

Fig. 3. The framework of the tree counting based on the Cascade Mask R-CNN, and the attention modules added to the backbone of the ResNet.

Table 1
The detailed parameters of the Cascade Mask R-CNN Network.

	Backbone: ResNet50	Neck: FPN	Regional Proposal Network (RPN)
Stages	4	5	–
In channels	3	256, 512, 1024, 2048 for each stage	256
Out channels	256, 512, 1024, 2048 for each stage	256	256
Anchor_ generator	–	–	Scale: 8 Ratio: [0.5, 1.0, 2.0] Step: [4, 8, 16, 32, 64]
Loss function	–	–	Classification: CrossEntropyLoss BBox Regression: GIoULoss
	ROI-head: Cascade Roi Head/RoIAlign	BBox-head	Mask Head
Stages	Cascade Stage: 3 The weight of each stage: [1, 0.5, 0.25]	–	4
In channels	256	256	256
Out channels	BBox prediction: $7 \times 7 \times 256$ Mask prediction: $14 \times 14 \times 256$	Fully connected layer output: 1024	256
Loss function	–	Classification: CrossEntropyLoss BBox Regression: GIoULoss	CrossEntropyLoss

$$\begin{cases} FP = DT - TP \\ FN = GT - TP \\ Precision = TP/DT \\ Recall = TP/GT \\ F1 = 2 \times Precision \times Recall/(Precision + Recall) \\ DR = DT/GT \end{cases} \quad (1)$$

where DT is the number of trees detected, and GT is the total number of trees labeled as ground truth in each plot.

We also used the coefficient of determination R^2 and root mean square error (RMSE). In this case, we compared the number of trees labeled tree and number of trees detected in each plot and calculated the R^2 and RMSE for the model assessment.

4. Results

4.1. Visual performance of tree counting

Fig. 4 shows the tree detection results for the different sceneries, with each being 1000 × 1000 pixels. The first row and second row in Fig. 4 show the trees in the parks by the river and the urban roadside trees, respectively. The plots had comparatively low canopy densities, and almost all of the trees were well detected by the four approaches, except YOLO, regardless of how large the crown was. Many of the shrubs in the top left corner in the first row were wrongly detected by the methods that contained attention modules.

The third row and forth row in Fig. 4 show the tree detection results in the residential area. Similar to the simple scenes in the first two rows in Fig. 4, good results were also achieved in the open residential area. The fifth row in Fig. 4 shows the trees scattered in the farmland. Again, the four approaches, except for YOLO, successfully distinguished the trees from the crops. In an orchard with relatively dense trees (last row in Fig. 4), the canopy was complicated because of the shadows and overlapping trees. The performance was slightly worse compared with the previous sceneries. The algorithm detected most of the trees, except for the trees with similar textures and indistinct crowns and some of the trees that blended into each other. Overall, the visual assessment of the tree detection results showed that most of the trees were well extracted, even the trees that slightly overlapped. The trees were somewhat

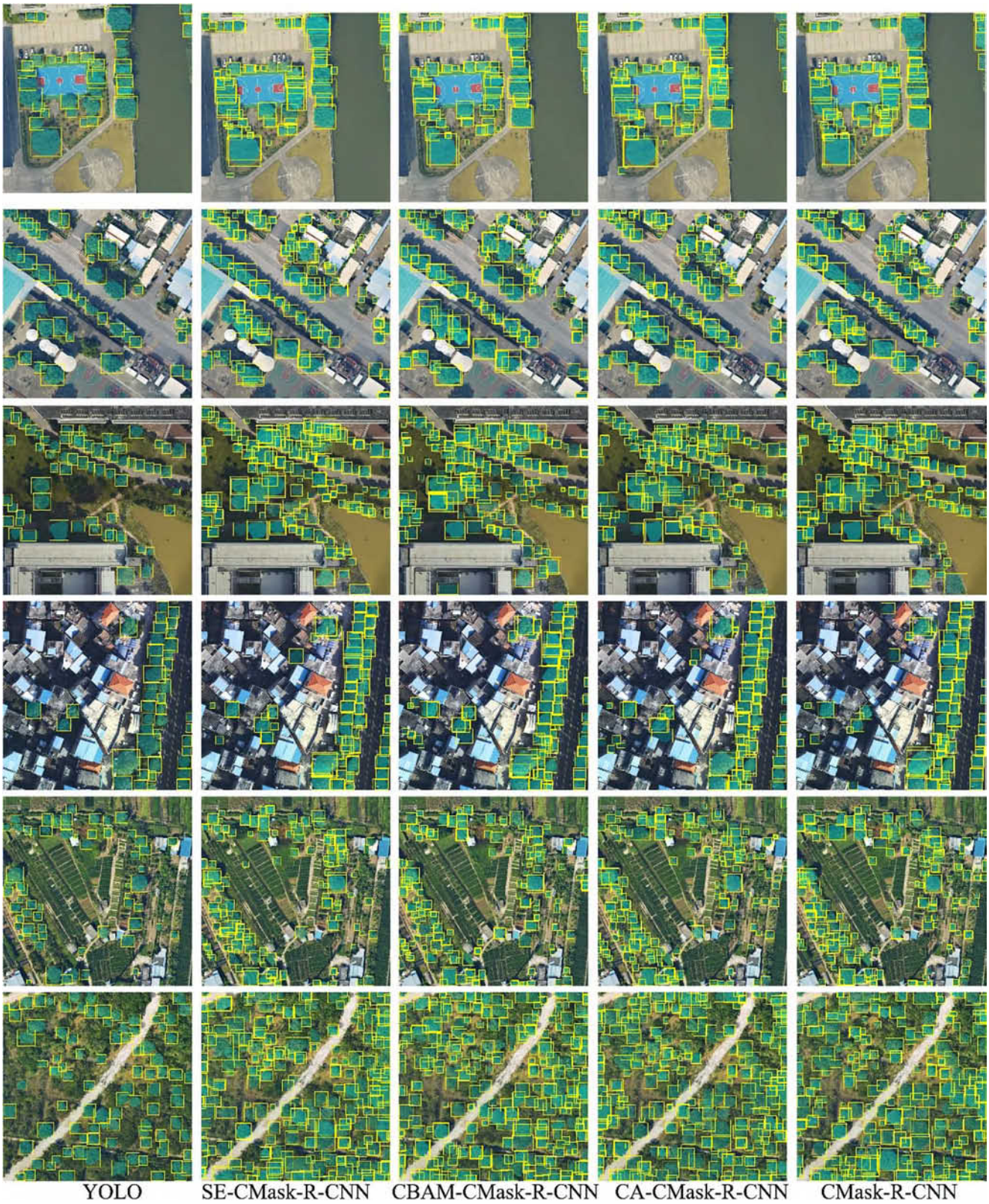

Fig. 4. Tree detection results for the different sceneries.

underestimated, however, as the dense canopies often formed a huge tree mass (last row in Fig. 4). Note that the trees in the high-density settlement area were well differentiated, including those hidden in the shadows of buildings. In terms of the methods, except YOLO, similar results were obtained visually. However, the CMask-R-CNN method had the best tree location detection results, that is, more trees were incorrectly detected by the methods containing attention modules than by the CMask-R-CNN compared with the ground truth.

4.2. Tree distribution in different land covers throughout Guangzhou

Based on the performances of the five methods, we used the tree-counting results of CMask-R-CNN for further analysis. We detected about 112 million trees with crowns larger than 1 m^2 throughout the entire city of Guangzhou (Table 2). Among the 11 districts, Conghua and Zengcheng had the most trees (i.e., 38.8 million and 28.8 million, respectively). They accounted for about 60% of the total trees in the city. This result was not surprising given that these two districts contain several large forest parks. Huadu and Baiyun contained the third and fourth largest number of trees, followed by Huangpu and Panyu. As the deputy center of Guangzhou, Nansha also had fewer trees than the other non-central region. As the three districts with the highest population densities, Haizhu, Liwan, and Yuexiu contained a total of about 2 million trees.

In addition to the number of trees, the tree density is a prominent component of the ecosystem structure and is important for modeling biological and biogeochemical processes (Crowther et al., 2015). Thus, we analyzed the tree density at the 1000 m scale (per 100 ha) and 30 m scale (per 900 m^2).

Fig. 5 shows the tree density map per 100 ha. Comparatively, the average canopy cover of the entire city was quite high (i.e., about 150 trees per hectare). The tree density distribution was mainly consistent with the forest coverage across the entire city. The tree density was high in the northern area of Guangzhou, followed by the central part. From the central-western part to the central-eastern part, the tree density increased. Although Yuexiu and Liwan had small numbers of trees, their tree densities were not the lowest. The tree density in the southern part, near the Pearl River estuary, was the lowest in Guangzhou. This is mainly because this area is a basic farmland conservation area. The built-up areas in the other non-central districts also had lower tree densities, such as the center of Conghua and northern part of Baiyun.

We also investigated how the trees were distributed in the different land cover types at a spatial resolution of 30 m (Fig. 6a). At this scale, similar tree density patterns were obtained per 100 ha, but more detailed information can be obtained. Fig. 6b shows the tree density in the urban land. Although the number of trees in the central area, such as the Yuexiu district, is lower than in Conghua and Zengcheng, the tree density in the urban land is similar throughout the entire city (i.e., about 15 trees per 900 m^2), indicating the good living environment. Among the central districts, Baiyun had the highest tree density in the urban land. Fig. 6c shows the tree density (an average of 32 trees per 900 m^2) in the forestland, which is concentrated in the central and northern parts of Guangzhou. We believe that it is underestimated in Fig. 6f. Of the two districts with rich forest resources, Conghua had a greater tree density than Zengcheng in the forestland. Fig. 6d shows the tree density in the grassland, which accounted for a small proportion and was scattered throughout the city. It almost had the lowest tree density compared with other land cover types. Fig. 6e shows the tree density in the farmland. All of the districts contained farmland, except for Yuexiu and Liwan. There were also trees in the farmland, with a density of about 20 trees per 900 m^2, and some of the trees were fruit trees.

Guangzhou is a subtropical city with large tree crowns. Fig. 7 shows the trees with crowns larger than 200 m^2 per 100 ha, and there are about 226,000 large trees. Based on visual inspection, these large trees were

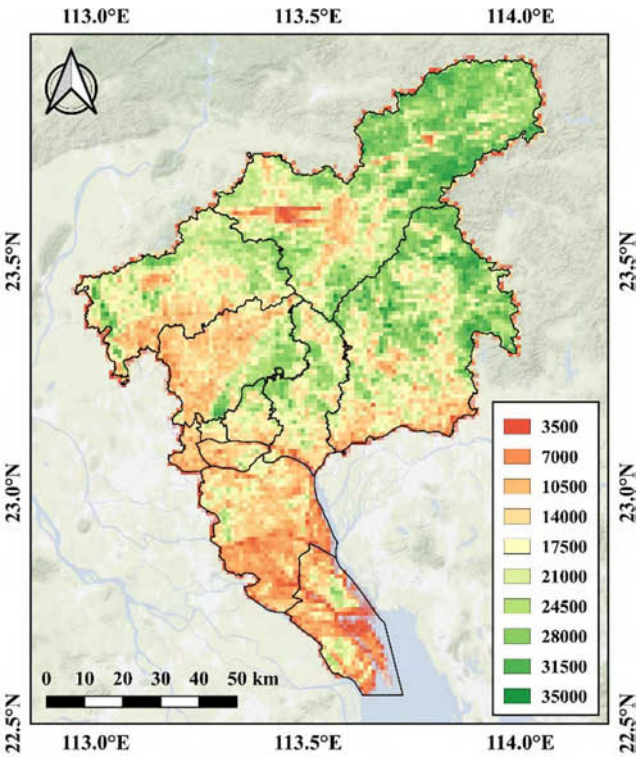

Fig. 5. Tree density map of Guangzhou per 100 ha.

located in all of the districts in Guangzhou, and most of the area with high densities of large trees were located in the central-western part. In the older parts of the city, such as Yuexiu, Liwan, and Tianhe, there were more than 275 large trees per 100 ha in some places. These trees are often banyan trees, and some have long histories and are ancient and famous trees with a high protection value. In addition, there were also a small number of areas with high densities of large trees in the northern part of Conghua, eastern part of Zengcheng, and southern part of Nansha.

We investigated the changes in the tree population in the past decade using bitemporal GlobeLand30 land cover maps. First, we obtained a map of the change in Guangzhou from 2010 to 2020. Then, we extracted the changes in the land use types: forestland-urban land, urban land-forestland, urban land-farmland, and urban land-grassland. Then, we overlapped the map of the changes (Fig. 8) and the tree density map to identify the tree changes caused by urbanization. The statistical results showed that about 200 km^2 changed from forest to urban land. Based on the current average tree density in the forestland, 4,416,459 trees were cut down and 2,694,652 trees were retained or replanted during the process of urbanization. About 689,005 trees were replanted through reforestation projects. Overall, the trees replanted are distributed in the forestland (about 38.96%), farmland (55.51%, including some fruit trees), and grassland (5.53%, including ornamental trees).

Table 2
The number of trees in each district of Guangzhou.

District	Number of trees	District	Number of trees
Baiyun	8,577,201	Nansha	3,499,431
Conghua	38,819,357	Panyu	6,837,273
Haizhu	984,068	Tianhe	2,010,680
Huadu	14,789,120	Yuexiu	382,079
Huangpu	7,329,621	Zengcheng	28,788,502
Liwan	569,556		

4.3. Tree-counting assessment

We produced a test dataset by randomly selecting tree samples. The test dataset contained 51 image patches (51 plots) with a total of 9021 trees, in which patch sizes of both 1000 × 1000 pixels and 1024 × 1024 pixels were used. We accounted for the different sceneries in the city in the tree sample selection, including urban land, parks, farmland, and forestland. Five instance segmentation methods were tested for the tree counting, and their accuracies are reported in Table 3. Among all of the methods, the CMask-R-CNN performed the best in terms of the error matrix, with an F1 of 82.56%. This index reflects the ability to detect the tree location. Although attention blocks have been shown to be effective

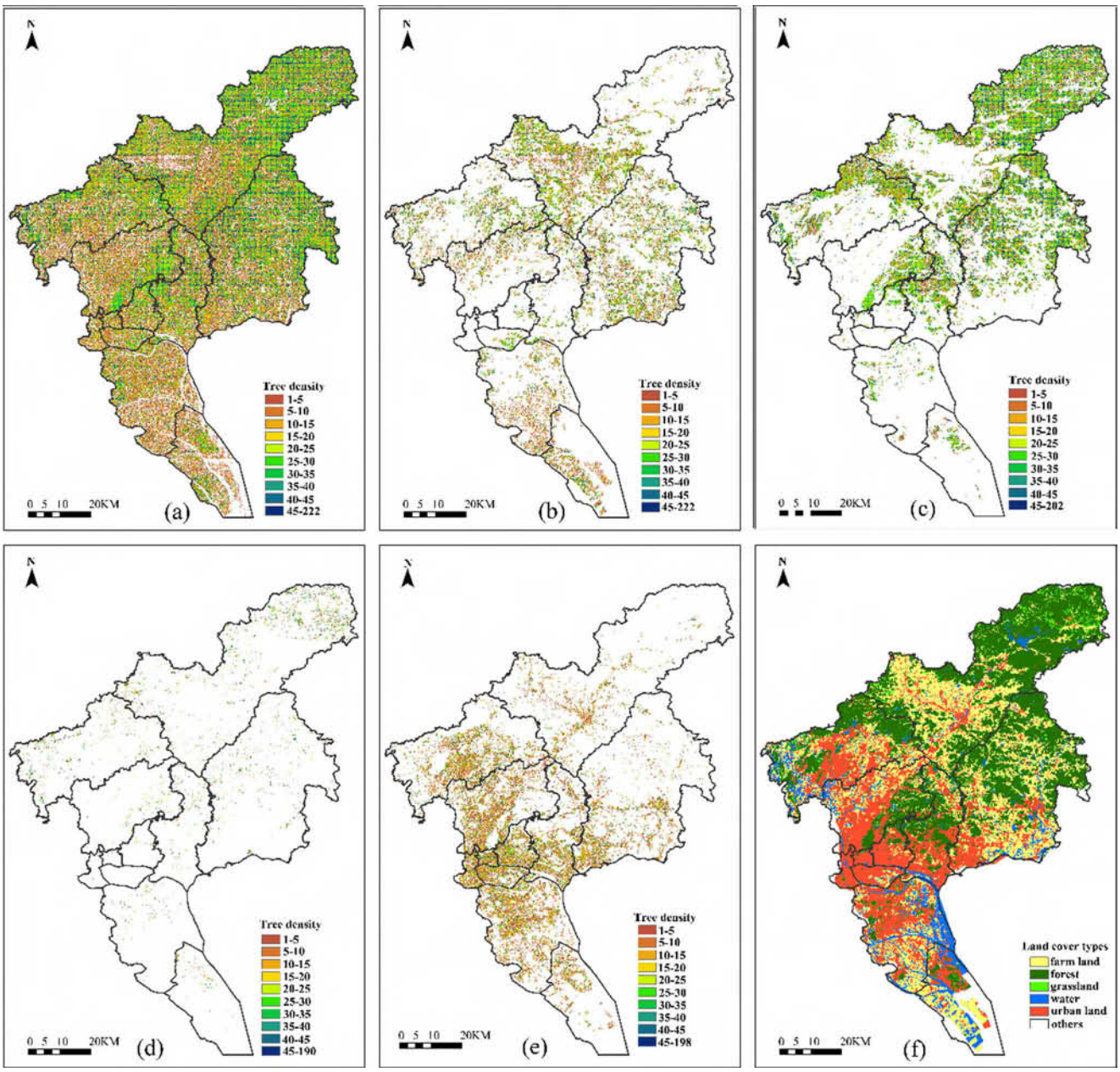

Fig. 6. Tree density distribution in different land cover types in Guangzhou at a spatial resolution of 30 m. (a) Tree density map of all of the land cover types, (b) Tree density map in the farmland, (c) Tree density map in forestland, (d) Tree density map in grassland, (e) Tree density map in urban land, and (f) land cover map of Guangzhou in 2020.

in many applications, the individual tree detection was not beneficial in strengthening the features derived from the backbone. In contrast to the detection of other objects, there was no obvious contrast between adjacent trees. The refinement of the features via an attention module may lose some useful features for tree crown delineation, and the information from different channel-wise or spatial-wise detection is important. Of the three types of attention modules, the CA-CMask-R-CNN achieved the best accuracy, indicating that it was effective at extracting features in both the X and Y directions.

Although the CBAM-CMask-R-CNN detected the closest number of trees to the ground truth, with a detection rate of 93.49%, many of the large trees were overestimated (Fig. 4). We also assessed the different methods using the R^2 and RMSE values (Fig. 9). The number of trees in each plot varied from dozens to more than 400 according to the different sceneries. The results were similar based on the metrics of the error matrix. The CMask R-CNN performed the best, with an R^2 of 88.32 and an RMSE of 47.34. Using the CMask R-CNN, 8249 trees were detected. Table S1 shows the accuracies of each plot with different canopy densities. The precision of all of the test plots is 86.42%. The metric of recall was worse than the precision, which indicated that the trees were often underestimated during the detection process. The most probable reason for this underestimation is that the detection method incorrectly identified overlapping canopies for one tree in the most challenging plot (last row in Fig. 4). The F1-score in Table S1 shows the weighted combination of the precision and recall, and the F1-score was 0.8256 for all of the test plots.

5. Discussion

According to the results presented thus far, we found that the tree-counting results varied for the different scenes. As is shown in Fig. 4, the methods detected the trees in the non-forest areas well. It performed

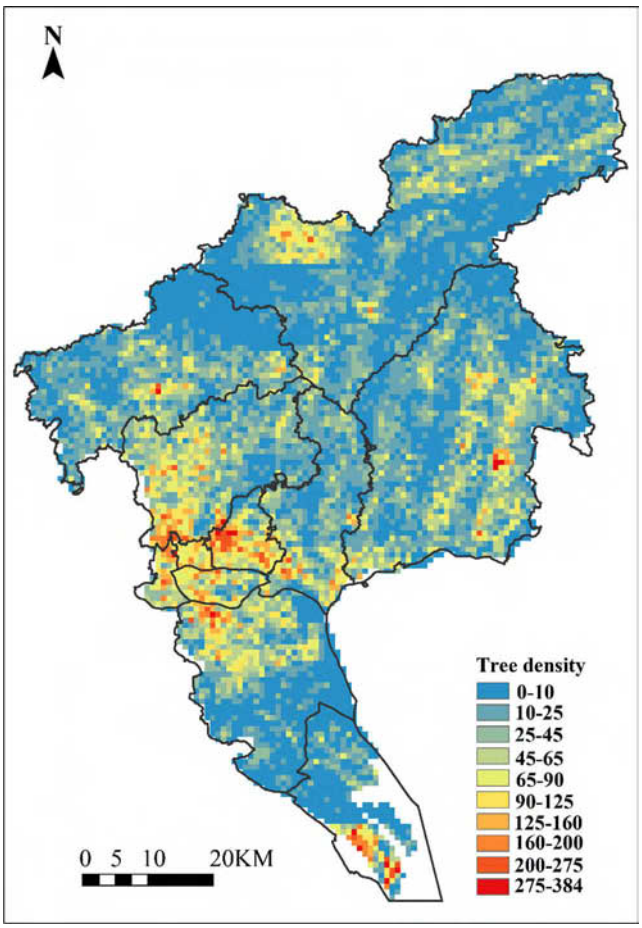

Fig. 7. Number of trees with crowns larger than 200 m^2 per 100 ha.

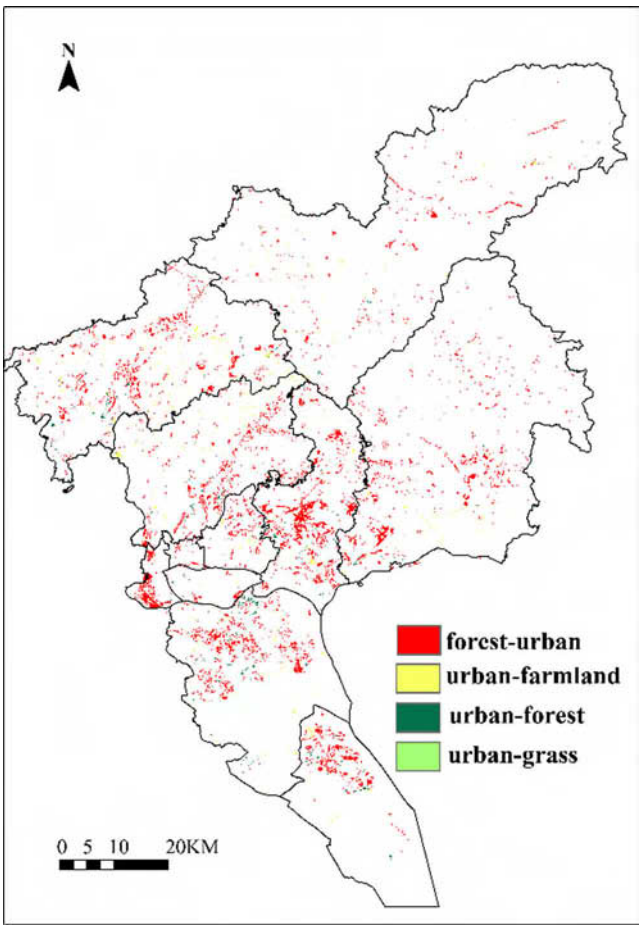

Fig. 8. Map of the land cover changes in Guangzhou showing four types of change.

well in the scenes with large contrast, such as water and impervious surfaces. Unexpectedly, it also performed well in the cropland and could distinguish scattered trees from crops. Unfortunately, the trees were underestimated in the densely distributed forest. In the dense forest, the results were often underestimated because some of the tree crowns connected with each other heavily and did not exhibit clear boundaries. For the trees with large crowns, however, the methods overestimated the number of trees because of the shadows of the large crowns. Based on all of the metrics, the CMask-R-CNN method had the best performance. Although the CBAM-CMask-R-CNN method achieved the highest tree detection rate, its F1 was lower than those of the other methods, except for YOLO, indicating that it incorrectly detected more trees than the CMask-R-CNN. The reason for this may be that the feature refinement through the attention modules may have lost some useful features for tree crown delineation. In addition, the one-step YOLO method lost the details of the high-resolution images and missed many trees.

In addition to the model's ability, the limitation of the data source was another important reason for accurate tree counting. The imaging conditions may affect the inclination of trees, the shadows of high buildings, and the ground object spectrum. The spectral resolution of the R-G-B images used was also limited. All of these factors influenced the individual tree detection. To improve the accuracy in the forestland, we would like to incorporate other data sources, such as hyperspectral images and LiDAR point clouds. Moreover, we will develop deep-learning networks for the combination of images and three-dimensional point cloud data to achieve better individual tree detection.

Although we accounted for different number of sceneries in the selection of the training samples across the entire city, the labeling of the trees in the forest area was very difficult because the tree crowns sometimes could not be recognized with the human eye. In addition, some were located deep in the primary forest, which was very difficult to visit. This also influenced the model's performance, especially in the areas of dense forest. Accordingly, it is necessary to label reliable tree crowns in dense areas to achieve better training of the deep instance segmentation network in the future.

Mapping the tree density is meaningful work for foresters, city managers, and scientists. Few studies have been conducted on this issue, and Crowther et al. (2015) is the only study to have mapped the tree density at a global scale. The spatial resolution of their work was 1 km^2, however, which was lower than that used for regional applications. Mapping the tree density at a finer scale, such as 30 m, is a good idea because it can clearly specify the detailed tree distribution in the area and also can be associated with 30 m resolution land cover maps. Therefore, many other fine-scale analyses can be achieved. In this study, we mapped the tree density at both the 1000 m and 30 m scales. The 1000 m scale was a moderate scale for investigating how the trees were distributed. The results of the 30 m scale analysis showed the tree density in the urban land, farmland, and grassland, which clarified the vegetation components of the land cover. Accordingly, we could infer the urban living environment, whether the farmland included an orchard, and whether there were landscape trees in the grassland. Moreover, this method provided an accurate underlying surface for land surface process modeling.

6. Conclusions

In this study, the number of trees in the subtropical mega city of Guangzhou in southern China was counted. We used an end-to-end tree-counting framework for the regional-scale tree detection. Based on VHR

Table 3
The accuracy of the tree detection results obtained using five instance segmentation methods.

	DT	GT	TP	FP	FN	Precision(%)	Recall(%)	F1 (%)	DR (%)
YOLO	3236	9021	3055	180	5994	94.41	33.87	49.85	35.87
CMask-R-CNN	8249	9021	7129	1102	1922	86.42	79.03	82.56	91.44
SE-CMask-R-CNN	7936	9021	6189	1732	2848	77.99	68.61	72.99	87.97
CA-CMask-R-CNN	8214	9021	6291	1915	2745	76.59	69.74	73.00	91.05
CBAM-CMask-R-CNN	8434	9021	6322	2101	2711	74.96	70.08	72.44	93.49

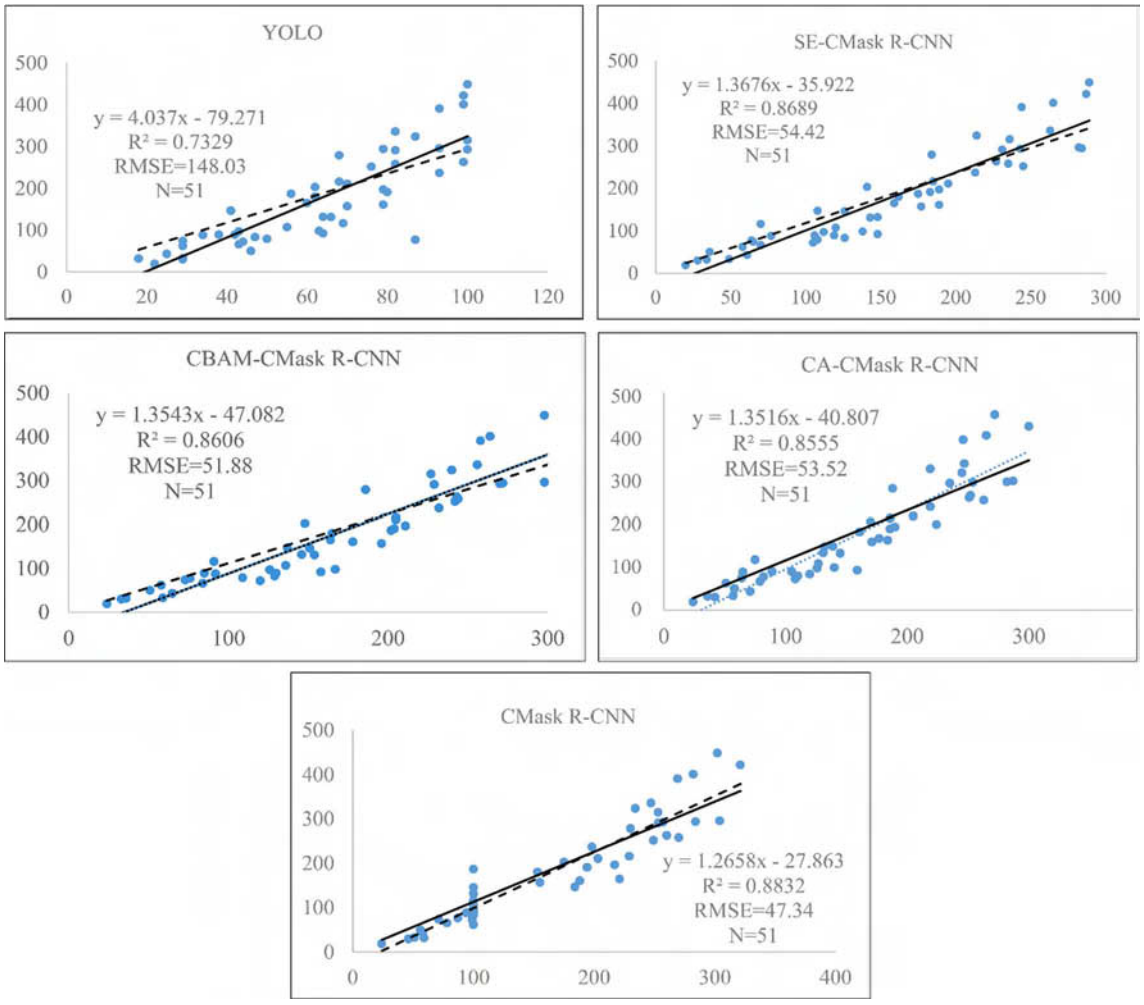

Fig. 9. The R^2 and RMSE values of the test dataset.

images, each tree crown was delineated. The framework allowed the input of image patch of any size within the memory of the GPU, and we used two sizes of 1000 × 1000 pixels and 1024 × 1024 pixels. The accuracies of five deep networks were assessed, among which the CMask-R-CNN performed the best. The experimental results showed that attention modules may not work well because the boundaries between adjacent trees are not always clear. Using the CMask-R-CNN, about 112 million trees were detected in Guangzhou, with an R^2 of 0.8832, an F1-score of 0.8256, and a DR of 0.9144. The accuracy assessment also revealed that the number of trees was somewhat underestimated. Moreover, we analyzed the tree density at the 30 m and 1000 m scales. Guangzhou contained 150 trees per hectare, which is a high canopy cover. Although urbanization took place in the city over the past decade, about 37% of the trees were retained or replanted in the newly added urban land during the process of urbanization. For the entire city, the tree density in the urban land was about 15 trees per 900 m^2, indicating a good living environment. The method developed in this study provides a flexible means of counting trees on a large scale without manual operation based on VHR imagery. This study not only revealed the number of trees but also provided the tree density at different scales, which is a prominent component of the ecosystem structure. The model code and the study materials will be made public soon.

CRediT authorship contribution statement

Ying Sun: Conceptualization, Methodology, Software, Writing – original draft. **Ziming Li:** Data curation, Software. **Huagui He:** Data curation, Visualization. **Xinchang Zhang:** Supervision. **Qinchuan Xin:** Writing – review & editing, Supervision.

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.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (grant nos. 42171308, 41801351), the Natural Science Foundation of Guangdong Province (grant no. 2021A1515011429), and the National Key R&D Program of China (grant nos. 2017YFA0604300 and 2017YFA0604400). We thank the anonymous reviewers for their constructive comments.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2021.102662.

References

Ammar, A., Koubaa, A., Benjdira, B., 2021. Deep-Learning-based Automated Palm Tree Counting and Geolocation in Large Farms from Aerial Geotagged Images. Agronomy 11, 1458.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495.
Brandt, M., Tucker, C.J., Kariryaa, A., et al., 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. Nature 587, 78–82.
Chan, T.-H., Jia, K., Gao, S., et al., 2015. PCANet: A simple deep learning baseline for image classification? IEEE Trans. Image Process. 24, 5017–5032.
Chen K., Pang, J., Wang, J., et al. 2019a. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974–4983.
Chen, L.-C., Papandreou, G., Kokkinos, I., et al., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.
Chen, Q., Baldocchi, D., Gong, P., et al., 2006. Isolating individual trees in a savanna woodland using small footprint lidar data. Photogramm. Eng. Remote Sens. 72, 923–932.
Chen, Z., Fan, W., Zhong, B., et al., 2019b. Corse-to-fine road extraction based on local Dirichlet mixture models and multiscale-high-order deep learning. IEEE Trans. Intell. Transp. Syst. 21, 4283–4293.
Chen, Z., Wang, C., Li, J., et al., 2021a. Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 100, 102341.
Chen, Z., Wang, C., Li, J., et al., 2021b. Reconstruction bias U-Net for road extraction from optical remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2284–2294.
Crowther, T.W., Glick, H.B., Covey, K.R., et al., 2015. Mapping tree density at a global scale. Nature 525, 201–205.
Crowther, T.W., Maynard, D.S., Leff, J.W., et al., 2014. Predicting the responsiveness of soil biodiversity to deforestation: a cross-biome study. Glob. Change Biol. 20, 2983–2994.
Duinker, P., Ordóñez, C., Steenberg, J., Miller, K., Toni, S., Nitoslawski, S., 2015. Trees in Canadian cities: indispensable life form for urban sustainability. Sustainability 7 (6), 7379–7396.
Duncanson, L., Cook, B., Hurtt, G., et al., 2014. An efficient, multi-layered crown delineation algorithm for mapping individual tree structure across multiple ecosystems. Remote Sens. Environ. 154, 378–386.
Erikson, M., 2003. Segmentation of individual tree crowns in colour aerial photographs using region growing supported by fuzzy rules. Can. J. For. Res. 33, 1557–1563.
Girshick, R., Donahue, J., Darrell, T., et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.
Hansen, M.C., Potapov, P.V., Moore, R., et al., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853.
He K., Gkioxari, G., Dollár, P., et al. 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.
He, K., Zhang, X., Ren, S., et al., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1904–1916.
He K., Zhang, X., Ren, S., et al., 2016. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.
Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Hou Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.
Khan, S., Gupta, P.K., 2018. Comparitive study of tree counting algorithms in dense and sparse vegetative regions. In: Int. Arch. Photogram, Remote Sensing Spatial Inform. Sci, pp. 801–808.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inform. Process. Syst. 25, 1097–1105.
Li, W., Guo, Q., Jakubowski, M.K., et al., 2012. A new method for segmenting individual trees from the lidar point cloud. Photogramm. Eng. Remote Sens. 78, 75–84.
Lin, T.-Y., Dollár, P., Girshick, R., et al., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Norzaki, N., Tahar, K.N., 2019. A comparative study of template matching, ISO cluster segmentation, and tree canopy segmentation for homogeneous tree counting. Int. J. Remote Sens. 40, 7477–7499.
Ocer, N.E., Kaplan, G., Erdem, F., et al., 2020. Tree extraction from multi-scale UAV images using Mask R-CNN with FPN. Remote Sensing Lett. 11, 847–856.
Osco L.P., de Arruda, M.d.S., Junior, J.M., et al. 2020. A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. ISPRS J. Photogram. Remote Sensing 160, 97–106.
Pan, Y., Birdsey, R.A., Fang, J., et al., 2011. A large and persistent carbon sink in the world's forests. Science 333, 988–993.
Pfeifer, M., Disney, M., Quaife, T., et al., 2012. Terrestrial ecosystems from space: a review of earth observation products for macroecology applications. Glob. Ecol. Biogeogr. 21, 603–624.
Qiu, L., Jing, L., Hu, B., et al., 2020. A new individual tree crown delineation method for high resolution multispectral imagery. Remote Sensing 12, 585.
Ren, S., He, K., Girshick, R., et al., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. 28, 91–99.
Rizeei, H.M., Shafri, H.Z.M., Mohamoud, M.A., Pradhan, B., Kalantar, B., 2018. Oil palm counting and age estimation from worldView-3 imagery and LiDAR data using an integrated OBIA height model and regression analysis. Journal of Sensors 2018, 1–13.
Romero-Lankao, P., Bulkeley, H., Pelling, M., et al., 2018. Urban transformative potential in a changing climate. Nat. Clim. Change 8, 754–756.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Salamí, E., Gallardo, A., Skorobogatov, G., et al., 2019. On-the-fly olive tree counting using a UAS and cloud services. Remote Sensing 11, 316.
Santoro, F., Tarantino, E., Figorito, B., et al., 2013. A tree counting algorithm for precision agriculture tasks. Int. J. Digital Earth 6, 94–102.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Stereńczak, K., Kraszewski, B., Mielcarek, M., et al., 2020. Mapping individual trees with airborne laser scanning data in an European lowland forest using a self-calibration algorithm. Int. J. Appl. Earth Obs. Geoinf. 93, 102191.
Tong, P., Han, P., Li, S., et al., 2021. Counting trees with point-wise supervised segmentation network. Eng. Appl. Artif. Intell. 100, 104172.
Tuanmu, M.N., Jetz, W., 2014. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Glob. Ecol. Biogeogr. 23, 1031–1045.
Vibha, L., Shenoy, P.D., Venugopal, K., et al., 2009. Robust technique for segmentation and counting of trees from remotely sensed data. In: 2009 IEEE International Advance Computing Conference. IEEE, pp. 1437–1442.
Wagner, F.H., Ferreira, M.P., Sanchez, A., et al., 2018. Individual tree crown delineation in a highly diverse tropical forest using very high resolution satellite images. ISPRS J. Photogramm. Remote Sens. 145, 362–377.
Weinstein, B.G., Marconi, S., Aubry-Kientz, M., et al., 2020. DeepForest: A Python package for RGB deep learning tree crown delineation. Methods Ecol. Evol. 11, 1743–1751.
Woo, S., Park, J., Lee, J.-Y., et al., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.
Wu, X., Shen, X., Cao, L., et al., 2019. Assessment of individual tree detection and canopy cover estimation using unmanned aerial vehicle based light detection and ranging (UAV-LiDAR) data in planted forests. Remote Sensing 11, 908.
Yan, W., Guan, H., Cao, L., et al., 2020. A self-adaptive mean shift tree-segmentation method using UAV LiDAR data. Remote Sensing 12, 515.
Yao, L., Liu, T., Qin, J., et al., 2021. Tree counting with high spatial-resolution satellite imagery based on deep neural networks. Ecol. Ind. 125, 107591.
Yin, D., Wang, L., 2016. How to assess the accuracy of the individual tree-based forest inventory derived from remotely sensed data: a review. Int. J. Remote Sens. 37, 4521–4553.
Zheng, J., Fu, H., Li, W., et al., 2020. Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network. ISPRS J. Photogramm. Remote Sens. 167, 154–177.
Zhou, J., Zhou, G., Li, Y., 2017. Above-Ground biomass estimation of larch based on terrestrial laser scanning data, 2017. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 6209–6212.
Zhu, Y., Newsam, S., 2017. Densenet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 790–794.