IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 5, 2022, accepted April 27, 2022, date of publication May 2, 2022, date of current version May 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171585

Individual Tree Detection Based on High-Resolution RGB Images for Urban Forestry Applications

LISHUO ZHANG[1,2], **HONG LIN**[2], **AND FENG WANG**[2]

[1] School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China
[2] Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 510060, China

Corresponding author: Lishuo Zhang (zhanglsh26@mail.sysu.edu.cn)

This work was supported in part by the Guangzhou Postdoctoral Research Project, in part by the Key Area Research and Development Program of Guangdong Province under Grant 2020B0101130009, and in part by the Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning under Grant 2020B121202019.

ABSTRACT Urban forests play an important role in urban ecosystems. They can not only beautify the urban environment but also help protect biodiversity and maintain ecological balance. Effective urban forest management is a basic requirement to ensure sustainable development. Traditional urban forest management usually requires the investment of a lot of materials and labor to conduct field research. RGB high-resolution aerial images have emerged as an efficient source of data for use in the detection and mapping of individual trees in urban areas. In recent years, there has been impressive progress in the field of deep learning methods for use in object detection. Semi-supervised learning is an effective way to deal with the problem that deep learning requires a large amount of labeled data. In this paper, we proposed an improved faster region-based convolutional neural network (Faster R-CNN) with Swin transformer method. Based on existing datasets, the model was trained and then transferred to new datasets. The method was evaluated within three distinct urban areas: a green space, a residential area and a suburban area. The experimental results indicate that our method achieved higher performance than other Faster R-CNN models. This method provides a reference in automated individual tree detection based on high-resolution images in urban areas for urban forestry managers.

INDEX TERMS Individual tree detection, swin transformer, faster R-CNN, urban forestry.

I. INTRODUCTION

Urban cities are built as human settlements and are the main areas of activity; meanwhile, they significantly contribute to climate change, and similarly, levels of vehicle exhaust emissions, incineration firing and industrial waste gas emissions. These threats have severely negative influences on the mental and physical health of those who live in urban cities. Urban trees warrant and support a range of vital social and environmental services to improve the air quality of cities in terms of air pollutants and particulate matter, increase the resilience of habitats and lower the urban heat islands effect [1], [2]. A forest resource survey is an indispensable and important link in urban forestry management. It can provide deep understanding of the status of forest resources and provide a reliable basis for forest resource management.

In previous studies, high-resolution images have been recommended for individual tree detection [3]–[10], high-resolution AGB estimation [11]. These studies verified the possibility of distinguishing urban forest canopies from other urban land covers with high accuracy. Jiao et.al [6] segmented trees and acquired locations and radii of the proposed cost function. In particular, the authors utilized shadows to calculate tree heights considering the image-taking conditions in terms of the sun angle and the time when each image was taken. Parmehr *et al.* [7] used the random forest model to achieve the detection of tree crown with overall accuracy of 79.3% over satellite imagery. Furthermore, in complicated urban areas, Ucar *et al.* [8] extracted woody vegetation by combining airborne imagery and Airborne Laser Scanning (ALS) data with an overall accuracy of 80%. Urban

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

environments are complex, which can be caused by cars, buildings, electronical lines and shadows. These hinder the automatic and efficient acquisition of data regarding tree canopies. However, the use of individual tree crown detection applications in land parcels can significantly support policy mechanisms [12].

Object detection aims to locate individual occurrences of a class (e.g., trees) within an image [13]. It is beneficial to account for carbon procedures at the stand, landscape and national levels [14]. In particular, this is crucial in urban areas, due to the fact that these images are heterogeneous and complex. Due to the fact that they are cost effective and easy to access, Aerial RGB images are widely applied, even though they lack three-dimensional information [15]. Additionally, there is spectral similarity between species in RGB scenes, which can be a hindrance for most automatic methods [16]. To avoid the object detection of RGB images from being labor intensive and cost intensive, these problems must be addressed.

Deep learning (DL) a field that is currently trending in machine learning and it focuses on fitting large models with millions of parameters for a variety of tasks [17]. DL often employs Convolutional Neural Networks (CNNs) to detect objects based on image classification and anchor box regression [18]. Faster R-CNN, which is a well-known object detection model, gives high-recall region proposals at low cost through Region Proposal Network (RPN), can significantly improve the efficiency of object detection. XIA *et al.* [23] developed an FPN-Faster R-CNN model, combining a feature pyramid network (FPN) and a Faster R-CNN. Santos *et al.* [19] chose Faster R-CNN, YOLOv3 and RetinaNet to evaluate their application performance in a forested urban area in Brazil. The authors explored the potential of DL in an experimental way. ZamboniThgeThe *et al.* [20] tested three advanced architectures: Faster R-CNN, RetinaNet and ATSS. They compared methods in two dimensions: quantitative and qualitative task completion, wherein ATSS performed better in a quantitate inspection and Faster R-CNN and RetinaNet were shown to have higher accuracy. For further development, the authors extended their study to investigate 21 novel deep learning methods; this provided a valuable reference for the application of deep learning [21]. Culman *et al.* [22] implemented RetinaNet to distinguish isolated and densely distributed date and canary palms with other Phoenix palms in a straightforward way. The results reported a mean average precision of 0.86. ROSLAN *et al.* [18] integrated a GAN based model and a RetinaNet model to detect individual tree crowns; the results showed excellent F1 scores. However, datasets are the foundation of deep learning, which need large amounts of labeled training data. Manual labeling is time consuming and labor intensive. LiDAR based individual tree detection was used to create sample data, and then, the Retinanet model was used to detect trees in the image [23], [24]. The use of existing datasets for deep learning tree detection is one motivation of this paper.

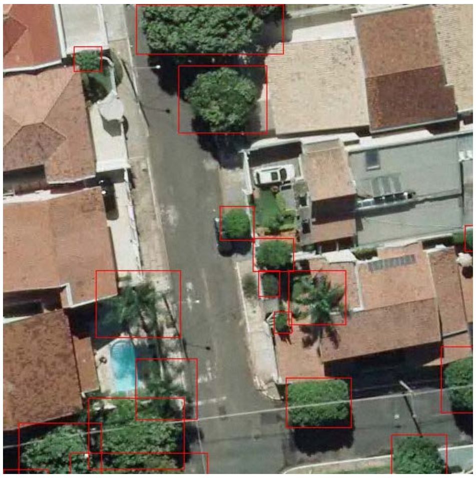


FIGURE 1. An annotated patch of Campo Grande dataset. The bounding boxes for each tree considered as ground truth are represented in red.

CNNs (VGG, Resnet34 and so on) were used to extract features in Faster R-CNN. However, a conventional feature extraction network encounters difficulty in producing abundant features due to the restricted size of the area in the input image. Dense and high-resolution predictions are required in these receptive fields. Previously, CNNs have not fully leveraged various feature maps from convolution or attention blocks conducive to object detection. Swin transformer [25] has a great capacity in sequence-based image modeling. Transformers are the first models that completely depend on self-attention to calculate input and output representations. The self-attention window improves feature extraction via local self-attention window calculation and cross-window connectivity. This hierarchical network can be used to make predictions at multiple feature scales.

In this paper, a transformer-based feature extraction network was introduced in Faster R-CNN to detect trees in urban areas. Based on existing datasets, transfer learning was used to retrain the model so that it could be used for predictions in new dataset. To the best of our knowledge, our study is the first time that Swin Transformer was used in tree detection in urban areas. The goals of our work were as follows:

1) To fine tune the model weights by appropriating Swin transformer as a backbone.

2) Based on the idea of transfer learning, to use labeled datasets to predict different types of datasets in different regions.

II. STUDY AREA AND DATA

Three primary sources of data were used in our experiments: one public and three private datasets.

A. CAMPO GRANDE DATASET

The Campo Grande dataset [21] includes two RGB high-resolution, airborne orthoimages with a ground sample distance (GSD) equal to 0.1m of the Campo Grande urban area, Mato Grosso do Sul state, Brazil. The orthoimages were

TABLE 1. Guangzhou dataset.

Field	AREA (KM ^{2})	Type	GSD(m)
Site1	0.18	Green space	0.1
Site2	0.57	Residential area	0.1
Site3	1.10	Suburb area	0.1

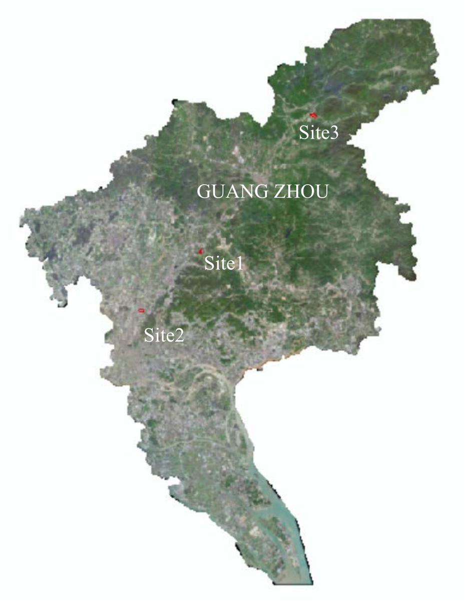


FIGURE 2. Location of study sites.

split into 220 non-overlapping patches of 512 × 512 pixels (Figure 1). A total of 3382 trees were identified as ground truth.

B. GUANGZHOU DATASET

Three study areas were selected to test the method. The three sites were located in Guangzhou, Guangdong province, China, as shown in Figure 2. The first study site is green space in the city, which covers 0.18km^2. The second study site is a habitation composed of buildings, a pond and few trees, and the last site is located in the suburbs, which covers 1.1km^2. These different study sites were selected because the trees included in these sites have different distribution characteristics, rendering it feasible to analyze the performance of the method for this study. Table 1 displays the three sites in this study. Ground Sample Distance (GSD) is 0.1m.

UAV based RGB imagery was acquired in April-October 2020 using a DJI Phantom 4 Pro (DJI Technology Co., Ltd., Shenzhen, China). The raw images acquired using the drone were processed using Agisoft Metashape v.1.5.5. The orthoimages with a GSD equal to 0.1m were used. Figure 3-5 show the orthoimages of Site 1, Site 2 and Site 3.

III. METHOD

A. FASTER R-CNN

Faster R-CNN, which is a two-stage detector, consists of three sections: 1) the feature extraction network; 2) the region proposal network (RPN); and 3) the region of interest (ROI) head. The feature extraction network was used to produce a feature map. By using a low-dimensional convolution layer, the RPN was implemented by scanning each region on the learned feature map, and then, multiple proposals on the feature map were predicted for each region. The proposals

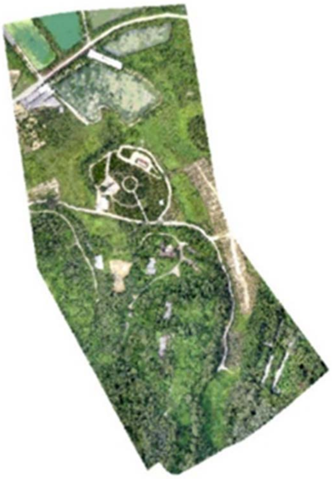


FIGURE 3. Orthoimage of site 3.

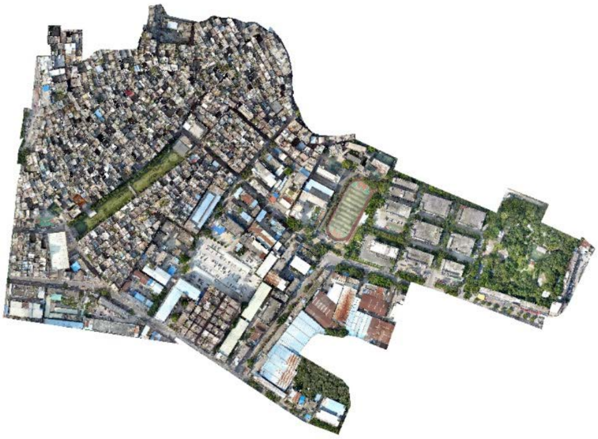


FIGURE 4. Orthoimage of site 2.

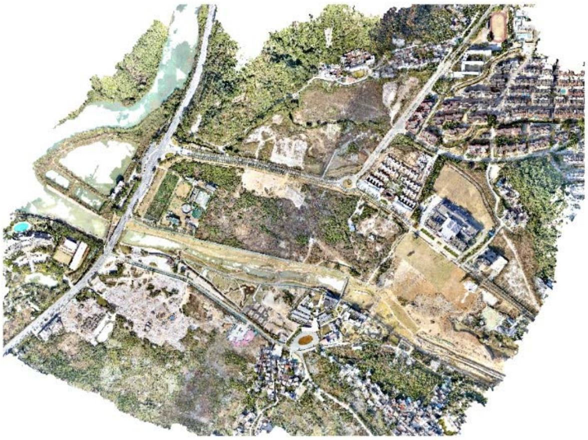


FIGURE 5. Orthoimage of site 1.

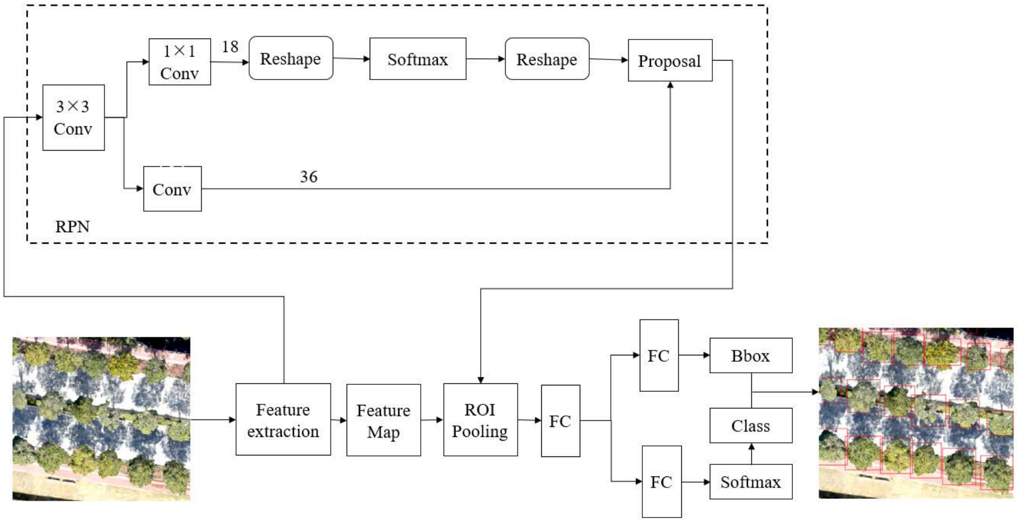
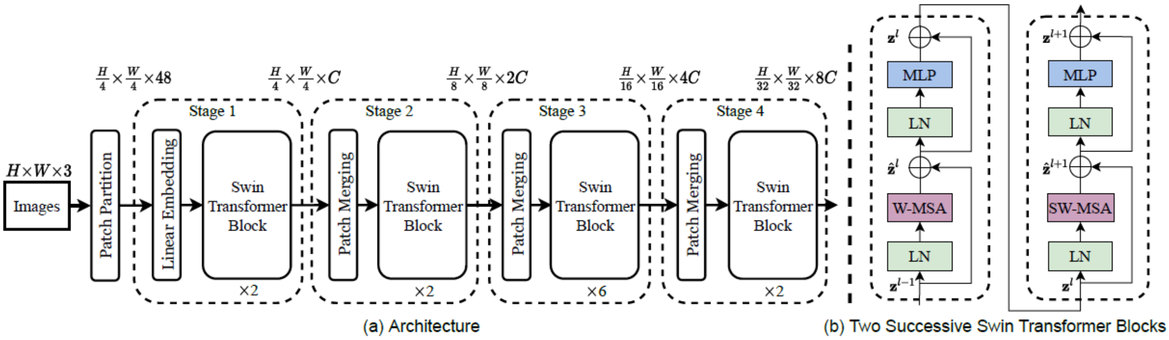


FIGURE 6. Faster R-CNN model.



(a) Architecture

(b) Two Successive Swin Transformer Blocks

FIGURE 7. (a) The architecture of a Swin transformer (Swin-T); (b) two successive Swin transformer blocks (W-MSA and SD W-MSA are window-based multi-head self-attention and window-based multi-head self-attention with spatial displacement, respectively [25].

were operated with ROI pooling to acquire the feature vectors. Subsequently, two fully connected layers (FCs) were adopted to predict the class and location of the proposals [26]. Figure 6 presents the traditional Faster R-CNN model.

As reported in [21], two-stage methods have higher performance in object detection. Here, firstly, regions that could have contained objects were filtered, and then, most negative regions were eliminated.

B. SWIN TRANSFORMER FEATURE EXTRACTION

As shown in Figure 7a, the input image was segmented into non-overlapping patches by a patch partition layer. Each patch was handled as a ''token'', and its feature could be regarded as a string of raw pixel values. The architecture of Swin transformer has four stages. Considering a H × W image, a token is a vector of an image patch with the size of 4 × 4. Linear embedding is applied on this token to map it in a vector with dimension C. In the architecture, stages 1-4 produce H/4 × W/4, H/8 × W/8, H/16× W/16 and H/32× W/32 tokens, respectively. Each stage consists of a patch merging block, a local perception block and some Swin transformer blocks. The detailed structure of the Swin transformer block is shown in Figure 7b. The block consists of window-based multi-head self-attention (W-MSA), shifted windows multi-head self-attention (SW-MSA) and multilayer perceptron (MLP). Inserting a layer norm (LN) layer in the middle makes the training more stable and uses a residual connection after each module.

C. IMPLEMENTATION

Considering that the feature extraction and utilization parameters of the basic network are relatively insufficient, this study drew on the intensive use of Swin transformer as a backbone to extract feature maps. To overcome the shortcomings of CNNs' poor ability to extract global information, we chose the Swin transformer as a basic backbone network to build a network model for individual tree detection in high-resolution RGB images.

The methodology work-flow chart is shown in Figure 8. Firstly, the network was trained by using Campo Grande dataset to detect the Guangzhou dataset. Subsequently, the predicted results with high confidence score were used to

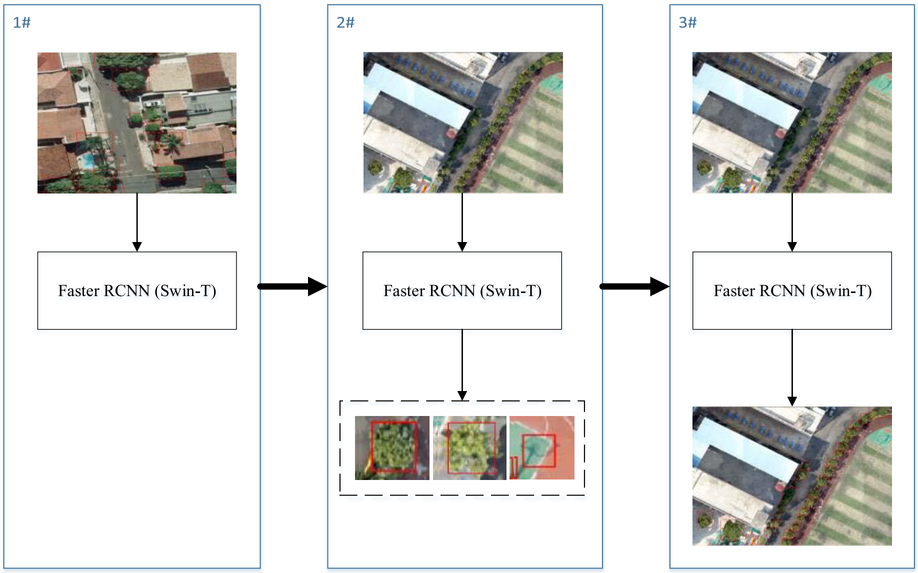


FIGURE 8. General processing chain: The detection network was trained by using Campo Grande dataset, Once trained, the network was applied to detect trees in Guangzhou dataset. The predicted results with high confidence score were used to retrain the detection model, Once retrained, the network was applied to detect trees.

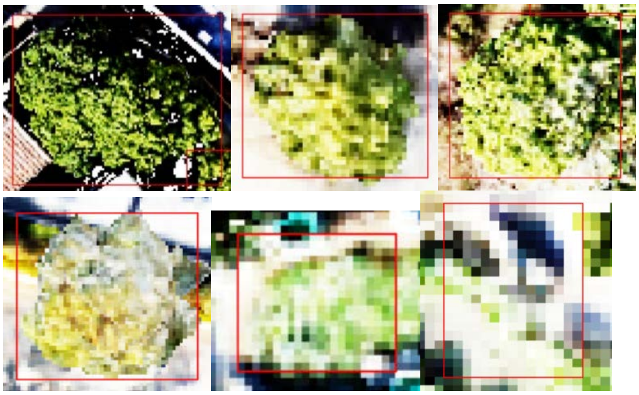


FIGURE 9. Results predicted by initial model: (a) score = 0.957, (b) score = 0.926, (c) score = 0.897, (d) score = 0.668, (e) score = 0.281, (d) score = 0.183.

retrain the detection model. The retrained network was applied to predict trees in the images not used for training. Finally, the accuracy metrics were computed on the predicted results.

Transfer learning is a technique of fine tuning a pre-trained model which can result in decently performing models for tasks with limited data. In this study, the Campo Grande dataset could be used to train our model. However, it was noted that without any fine tuning, the model did not perform very well in new datasets. Manual labeling is time consuming and labor intensive. There were no labeled data available to perform this fine tuning. Some noisy data labels were generated by this initial model which could be used to fine tune the model. Hence, a semi-supervised approach was used to fine tune the base model further. The labeled data were taken from the previously predicted results using the same model. However, the average confidence score was low. These data were filtered out and used as the retraining data. The true object should have had higher confidence scores, as shown in Figure 9.

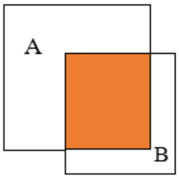


FIGURE 10. IoU.

A stochastic gradient descent optimizer with a momentum of 0.9 and weight decay of 0.0001 was applied. The initial learning rate was empirically set to 0.001[20]. We used Pytorch as the DL framework, and the compilation environment was Python 3.6.13 and Pytorch 1.8.0. The training and testing procedures were implemented with CUDA-compatible NVIDIA GPU (GeForce RTX 2080 super, 8 GB RAM).

D. PERFORMANCE EVALUATION

The Intersection over Union (IoU), which is the ratio between the union and the intersection of predicted box and ground truth box, was used to measure overlap, as shown in Figure 10 and (1). When a predicted box reaches a greater IoU than the threshold, the prediction is classified as true positive (TP). Otherwise, the prediction box is a false positive (FP). Furthermore, if a ground truth box is not detected, it is considered a false negative (FN). We calculated precision, recall, and F1 score at an IoU threshold of 0.5 for each image patch. Precision, recall and F1 Score were calculated according to (2)-(4).

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

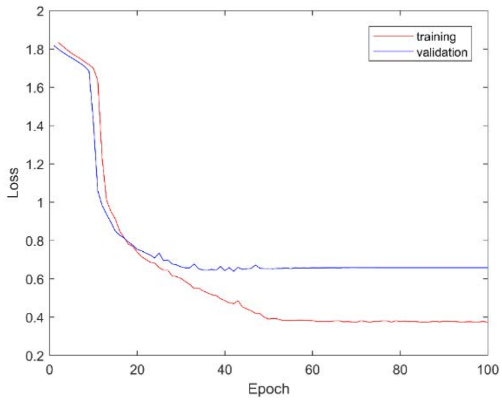


FIGURE 11. The training and validation loss curves of Faster RCNN (Swin-T).

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

IV. RESULTS

Here, we present the results of our experiments. We chose different types of experimental datasets to perform statistical analysis, which included the Campo Grande dataset and the Guangzhou dataset.

The results are organized in three sections: First, the Campo Grande dataset was trained using the proposed approach to provide an initial model; later, the model was applied to the Guangzhou dataset to predict tree canopies; after filtering the predicted result, the model was retrained.

A. CAMPO GRANDE DATASET

In this section, we discuss the performance of Faster R-CNN (Swin-T) and other models used in [21]. The training, validation, and testing sets comprising 60%, 20% and 20% of the available images, respectively.

Figure 11 illustrates the training and validation loss curves of 100 epochs. The training loss decreased rapidly after a few epochs, fluctuating in the following epochs, and then stabilized at the end. The stable trends indicated that the training epochs could meet the requirements.

Table 2 shows the performance for Faster R-CNN (Swin-T), Faster R-CNN (Resnet50) and some models in [19] with the average precision (AP50) and an IoU value of 0.5. The performance of Faster R-CNN (Swin-T) was shown to have similar precision to FSAF, which had the best performance with respect to AP50. Additionally, the Faster R-CNN (Swin-T) achieved higher performance than Faster R-CNN and RetinaNet.

B. GUANGZHOU DATASET

Here, we chose three kinds of land parcels covering residential, urban green space and suburban areas. Figures 12-14 show the tree detection achieved using Faster R-CNN (Swin-T) and Faster R-CNN (Resnet50) represented by a red box and a yellow box, respectively.

TABLE 2. Performance of the methods.

Model	AP_{50}
Faster R-CNN (Swin-T)	0.699
Faster R-CNN (Resnet50)	0.652
Faster R-CNN（Resnet101FPN）[21]	0.660
RetinaNet [21]	0.650
FSAF [21]	0.701

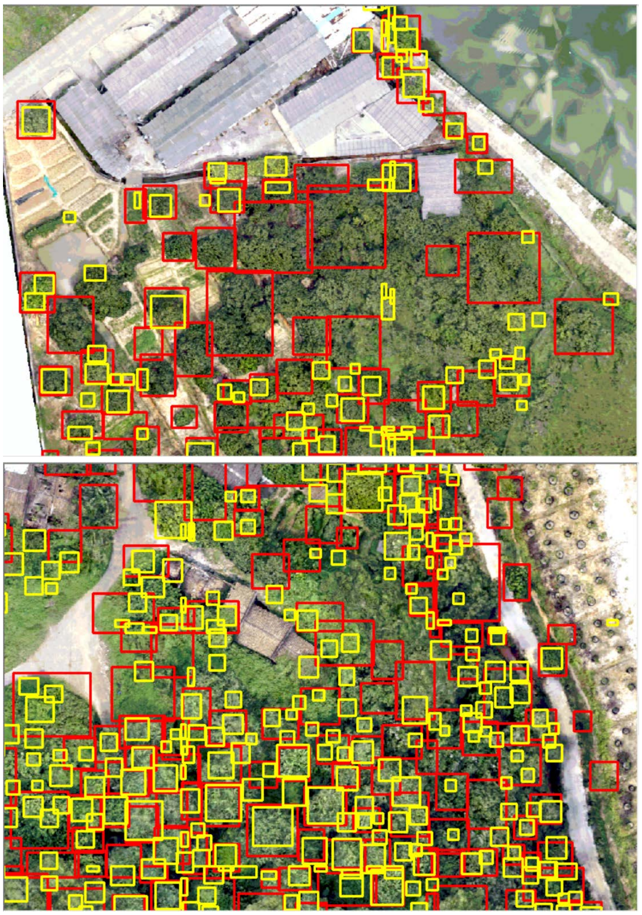


FIGURE 12. Predicted results regarding green space, red: Faster R-CNN (Swin-T); yellow: Faster R-CNN (Resnet50).

With regard to green space (Figure 12), as for smaller tree crowns and even medium-sized ones, Faster R-CNN (Swin-T) had good assertiveness. However, for larger crowns, we observed a decrease in the performance in areas where many trees overlapped with each other. Even so, Faster R-CNN (Resnet50) showed worse robustness under the same conditions.

As we can see in Figure 13, Faster R-CNN (Swin-T) has better assertiveness in areas with buildings and shadows. Correspondingly, there are instances of false detection, where buildings are recognized as trees using Faster R-CNN (Resnet50).

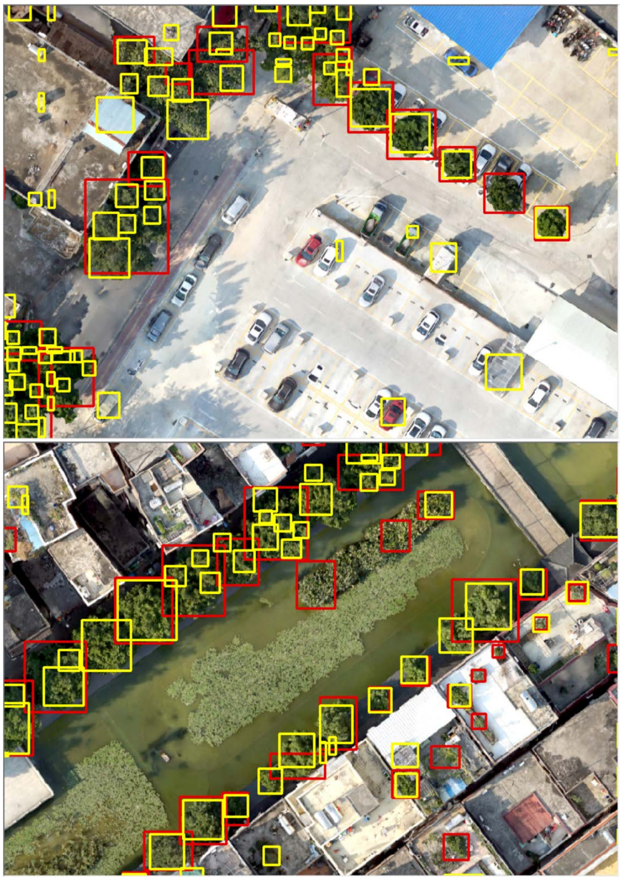


FIGURE 13. Predicted results regarding residential area, red: Faster R-CNN (Swin-T); yellow: Faster R-CNN (Resnet50).

Regarding suburban areas, despite missed detections of agglomeration or overlap of trees, Faster R-CNN (Swin-T) works well (Figure 14). As for less complex landscapes and the interference factor in suburban districts, synthetically, the proposed model can achieve excellent performance.

Table 3 presents tree detection counts of three overall land parcels for residential areas, green spaces and suburban districts. Additionally, Figures 15-17 show the application of the proposed model in different scenarios. The results demonstrate that our method can effectively geolocate trees and delineate tree distribution.

C. STREET TREE DETECTION

Street trees are an important part of urban ecological environments and have multiple functions such as dustproofing and noise reduction, shading. Therefore, it is necessary to monitor their growth status. The quick and cheap obtainment of the stock of street trees is an area of importance for urban foresters. Here, we tested the performance of methods with regard to the detection of street trees. We annotated a small amount of street trees, similar to Figure 1. As can be seen from Table 4, the F1 of the proposed method was 0.948, which was larger than the Faster R-CNN (Resnet50) score of 0.898; the proposed method was shown to have higher detection precision and recall, which led to higher F1 scores.

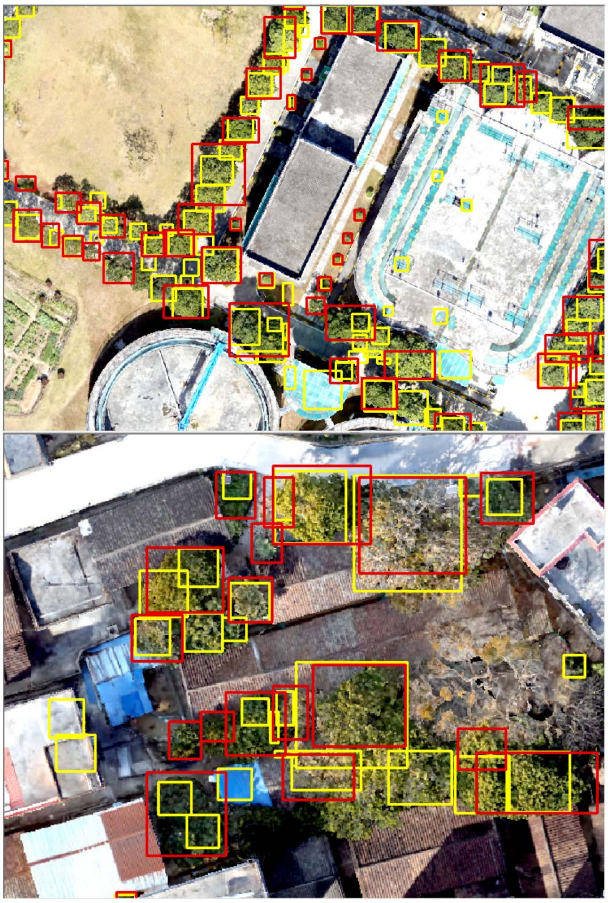


FIGURE 14. Predicted results regarding suburban area, red: Faster R-CNN (Swin-T); yellow: Faster R-CNN (Resnet50).

TABLE 3. Number of detected trees.

Field	Count
Site1	2241
Site2	1713
Site3	11459

Moreover, as shown in Figure 18, the Faster R-CNN (Swin-T) benefits from the attention mechanism and thus performs better than Faster R-CNN (Resnet50). The majority of street trees were detected by the Faster R-CNN (Swin-T) method. However, some false detections and missed detections were caused by Faster R-CNN(Resnet50). Thus, our method was shown to achieve much better detection performance than the comparison methods in terms of the detection of street trees.

D. COMPUTATIONAL COMPLEXITY

Table 5 shows the mean and standard deviation of the time for training(141 images) and validation(35 images). The image size is 512 × 512. The time is the average in seconds to execute the methods in an epoch. it becomes faster by using the proposed architecture.

Table 6 shows the mean and standard deviation of the time for tree detection. The time is the average in seconds

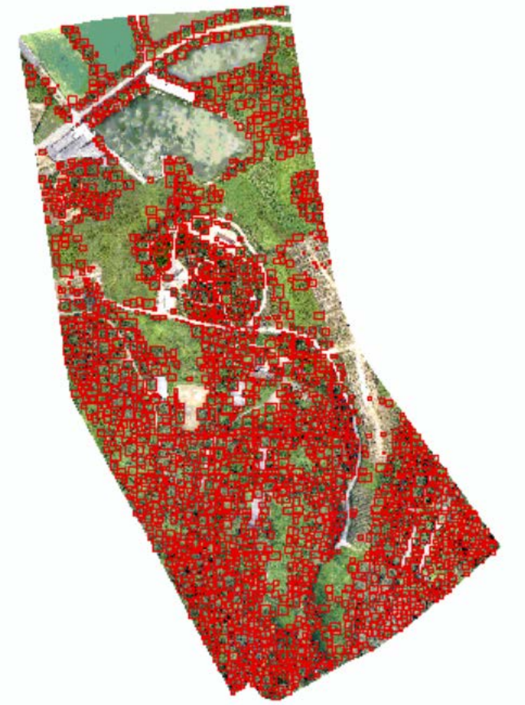


FIGURE 15. Predicted result of site 1.

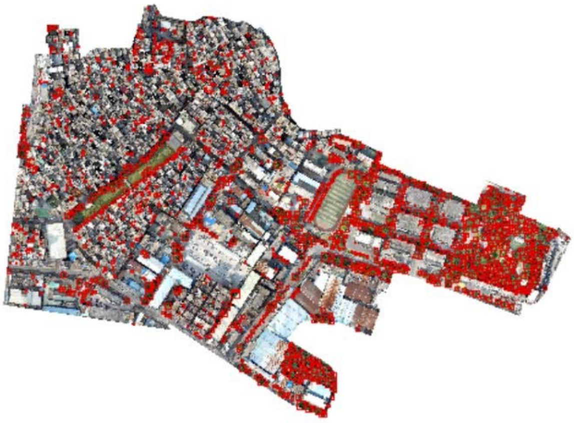


FIGURE 16. Predicted result of site 2.

to execute the methods on an image (512 × 512). it becomes faster by using the proposed architecture.

V. DISCUSSION

The results presented in the previous sections demonstrate that the proposed method could detect trees effectively in the study areas. An existing labeled dataset can be used to train deep learning models to recognize trees and transfer them to another dataset. However, as shown in the results, the method in this paper is not effective enough to handle an intersection of trees; an agglomeration of trees was detected as one tree. The method was tested in different urban land types: green space, suburban and residential areas. The proposed method worked well in suburban and residential areas in spite of some detection mistakes in the overlap between trees. The method used in this paper could detect some small trees on roofs. Similarly to other methods, the method in this paper was not effective in the detection of many dense forests, and the detection of high-density objects was also a problem [21].

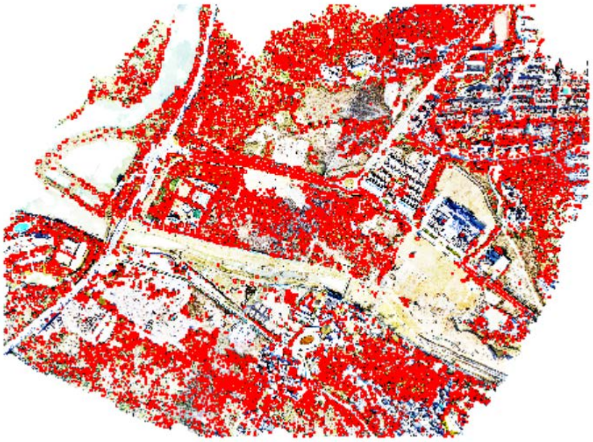


FIGURE 17. Predicted result of site 3.

TABLE 4. Performance of the methods.

Model	TP	FP	FN	P	R	F1
Faster R-CNN (Swin-T)	226	16	9	0.934	0.962	0.948
Faster R-CNN (Resnet50)	218	33	17	0.869	0.928	0.898

Comparably, the Faster R-CNN (Resnet50) model produced some false detections, as it recognized some houses, cars, etc., as trees, which rarely occurred in the proposed method. The Swin transformer has been used in remote sensing classification, object detection and instance segmentation [27]–[29]. As has been previously presented, the experimental results demonstrate the powerful ability of Swin transformer.

As urban scenes are more complex and heterogeneous, there were some hindrances in terms of tree delineation such as some overlap between objects and shadows and other situations. In these scenarios, LiDAR [30]–[32] or multispectral airborne LiDAR [33] can be integrated to enhance the detection accuracy by providing altitude information. The combination of LiDAR and imagery has recently become popular for use in urban forest inventories [34]. Zhang *et al.* [35] proposed a framework to segment individual trees in urban areas using airborne LiDAR and aerial images, and NDVI derived from hyperspectral images was used to separate vegetation points from point clouds, and then, individual trees were segmented from vegetation points. However, expensive costs and the complex processing flow limit the wide application of this technique.

Aval *et al.* [36] used airborne hyperspectral data and a DSM and a vector layer of roads derived from the OSM database to detect street trees, and the F1 score was shown to

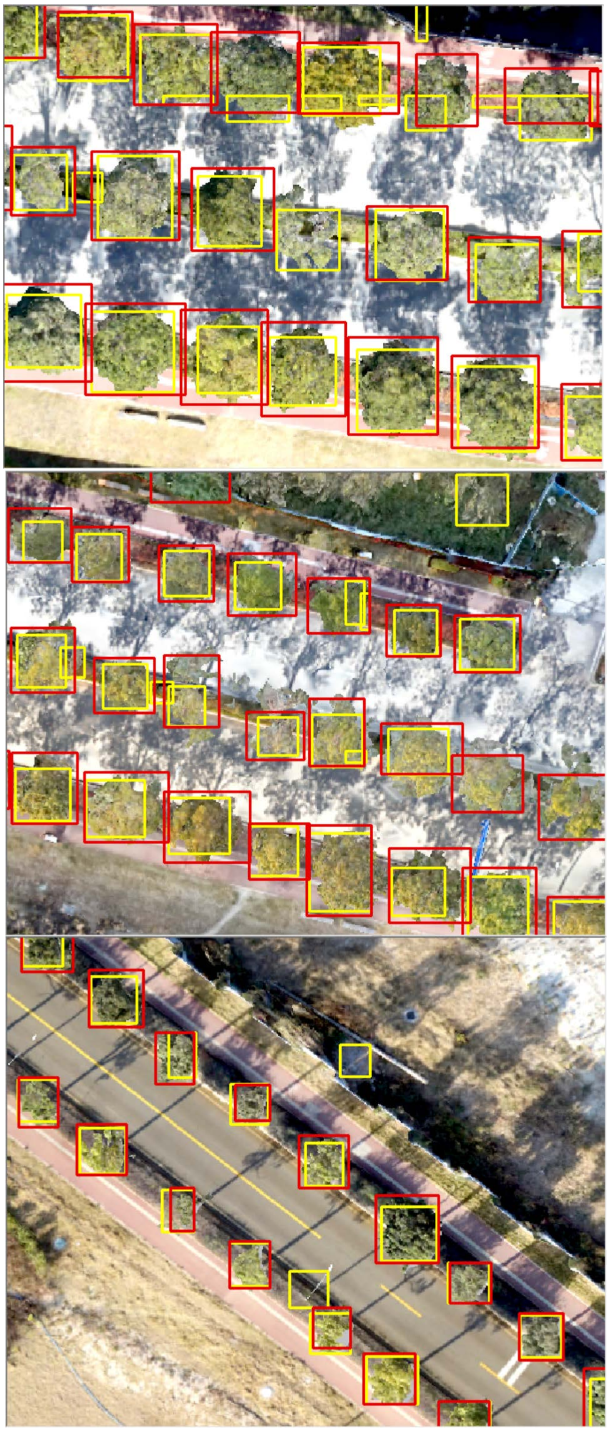


FIGURE 18. Street tree detection, red: Faster R-CNN (Swin-T); yellow: Faster R-CNN (Resnet50).

be 0.75-0.91. Our method could also detect street trees with high accuracy. However, it should be noted that in our dataset, the street trees were scattered and rarely overlapped with each other.

Even though the proposed method displayed good performance in experiments, our research aimed to detect single trees in an urban area. The detection of all of the trees in urban scenes is a considerably more challenging task. This area of research is still incipient, and it is worth further investigation regarding the most appropriate techniques and different types of data sources.

TABLE 5. Training time of the methods.

Model	Time (s)
Faster R-CNN (Swin-T)	35.760(±0.417)
Faster R-CNN (Resnet50)	40.527(±0.725)

TABLE 6. Computational cost of the methods.

Model	Time (s)
Faster R-CNN (Swin-T)	0.103(±0.062)
Faster R-CNN (Resnet50)	0.122(±0.067)

VI. CONCLUSION

We presented a deep learning method for the detection of individual trees in urban areas based on high-resolution RGB images. The developed model can achieve reutilization in other urban scenes. We explored the method using different datasets. The results show that the model achieved performance with an AP50 of 0.699 in the Campo Grande dataset. We also provided a qualitative analysis with regard to three land parcels of residential areas, suburbs and green areas, and the proposed architecture was shown to have better performance. This is especially true for street tree detection, where our method was shown to have better assertiveness; in particular, the F1 score reached 0.948 in this case.

Furthermore, the method used in this paper did not need to label data (or label a small amount of data for testing), which greatly reduced the burden of manual labeling and can be applied to other datasets. The final output of this method was a vector file, not just a target box, of which a spatial target was geolocated in the spatial dimension. This study provided valuable information for urban forestry practitioners and can be used in future works concerning the detection of individual trees.

REFERENCES

[1] C. Weber, "Ecosystem services provided by urban vegetation: A literature review," in *Urban Environment*. Dordrecht, The Netherlands: Springer, 2013, pp. 119–131.

[2] J. M. A. Duncan, B. Boruff, A. Saunders, Q. Sun, J. Hurley, and M. Amati, "Turning down the heat: An enhanced understanding of the relationship between urban vegetation and surface temperature at the city scale," *Sci. Total Environ.*, vol. 656, pp. 118–128, Mar. 2019.

[3] D. L. Torres, R. Q. Feitosa, P. N. Happ, L. E. C. La Rosa, J. M. Junior, J. Martins, P. O. Bressan, W. N. Gonçalves, and V. Liesenberg, "Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery," *Sensors*, vol. 20, no. 2, p. 563, Jan. 2020.

[4] J. A. McGee, S. D. Day, R. H. Wynne, and M. B. White, "Using geospatial tools to assess the urban tree canopy: Decision support for local governments," *J. Forestry*, vol. 110, no. 5, pp. 275–286, Jul. 2012.

[5] T. Erker, L. Wang, L. Lorentz, A. Stoltman, and P. A. Townsend, "A statewide urban tree canopy mapping method," *Remote Sens. Environ.*, vol. 229, pp. 148–158, Aug. 2019.

[6] J. Jiao and Z. Deng, "Individual building rooftop and tree crown segmentation from high-resolution urban aerial optical images," *J. Sensors*, vol. 2016, pp. 1–13, Jan. 2016.

[7] E. G. Parmehr, M. Amati, E. J. Taylor, and S. J. Livesley, "Estimation of urban tree canopy cover using random point sampling and remote sensing methods," *Urban Forestry Urban Greening*, vol. 20, pp. 160–171, Dec. 2016.

[8] Z. Ucar, P. Bettinger, K. Merry, R. Akbulut, and J. Siry, "Estimation of urban woody vegetation cover using multispectral imagery and LiDAR," *Urban Forestry Urban Greening*, vol. 29, pp. 248–260, Jan. 2018.

[9] Y. Lin, M. Jiang, Y. Yao, L. Zhang, and J. Lin, "Use of UAV oblique imaging for the detection of individual trees in residential environments," *Urban Forestry Urban Greening*, vol. 14, no. 2, pp. 404–412, 2015.

[10] X. Li, W. Y. Chen, G. Sanesi, and R. Lafortezza, "Remote sensing in urban forestry: Recent applications and future directions," *Remote Sens.*, vol. 11, no. 10, p. 1144, May 2019.

[11] K. Ghosal, S. Das Bhattacharya, and P. K. Paul, "Estimation of above-ground forest biomass in Himalayan region of West Bengal, India using IRS P6 LISS-IV data," *Arabian J. Geosci.*, vol. 15, no. 7, pp. 1–28, Apr. 2022.

[12] C. Ordóñez-Barona, J. Bush, J. Hurley, M. Amati, S. Juhola, S. Frank, M. Ritchie, C. Clark, A. English, K. Hertzog, M. Caffin, S. Watt, and S. J. Livesley, "International approaches to protecting and retaining trees on private urban land," *J. Environ. Manage.*, vol. 285, May 2021, Art. no. 112081.

[13] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, Mar. 2021.

[14] D. Pulido, J. Salas, M. Rös, K. Puettmann, and S. Karaman, "Assessment of tree detection methods in multispectral aerial images," *Remote Sens.*, vol. 12, no. 15, p. 2379, Jul. 2020.

[15] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks," *Remote Sens.*, vol. 11, no. 11, p. 1309, Jun. 2019.

[16] L. S. D. Arce, L. P. Osco, M. D. S. D. Arruda, D. E. G. Furuya, A. P. M. Ramos, C. Aoki, A. Pott, S. Fatholahi, J. Li, F. F. D. Araújo, W. N. Gonçalves, and J. M. Junior, "Mauritia flexuosa palm trees airborne mapping with deep convolutional neural network," *Sci. Rep.*, vol. 11, p. 19619, Oct. 2021.

[17] Y. Diez, S. Kentsch, M. Fukuda, M. L. L. Caceres, K. Moritake, and M. Cabezas, "Deep learning in forestry using UAV-acquired RGB data: A practical review," *Remote Sens.*, vol. 13, no. 14, p. 2837, Jul. 2021.

[18] Z. Roslan, Z. A. Long, and R. Ismail, "Individual tree crown detection using GAN and RetinaNet on tropical forest," in *Proc. 15th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2021, pp. 1–7.

[19] A. A. D. Santos, J. M. Junior, M. S. Araújo, D. R. Di Martini, E. C. Tetila, H. L. Siqueira, C. Aoki, A. Eltner, E. T. Matsubara, H. Pistori, R. Q. Feitosa, V. Liesenberg, and W. N. Gonçalves, "Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVs," *Sensors*, vol. 19, no. 16, p. 3595, Aug. 2019.

[20] P. A. P. ZamboniThgeThe, J. Marcato, G. T. Miyoshi, J. de Andrade Silva, J. Martins, and W. N. Goncalves, "Assessment of CNN-based methods for single tree detection on high-resolution RGB images in urban areas," presented at the IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2021.

[21] P. Zamboni, J. M. Junior, J. D. A. Silva, G. T. Miyoshi, E. T. Matsubara, K. Nogueira, and W. N. Gonçalves, "Benchmarking anchor-based and anchor-free state-of-the-art deep learning methods for individual tree detection in RGB high-resolution images," *Remote Sens.*, vol. 13, no. 13, p. 2482, Jun. 2021.

[22] M. Culman, S. Delalieux, and K. Van Tricht, "Individual palm tree detection using deep learning on RGB imagery to support tree inventory," *Remote Sens.*, vol. 12, no. 21, p. 3476, Oct. 2020.

[23] B. G. Weinstein, S. Marconi, S. A. Bohlman, A. Zare, and E. P. White, "Cross-site learning in deep learning RGB tree crown detection," *Ecol. Informat.*, vol. 56, Mar. 2020, Art. no. 101061.

[24] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks," *Remote Sens.*, vol. 11, no. 11, p. 1309, Jun. 2019.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[27] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, "An improved swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sens.*, vol. 13, no. 23, p. 4779, Nov. 2021.

[28] A. Jamali and M. Mahdianpari, "Swin transformer and deep convolutional neural networks for coastal wetland classification using Sentinel-1, Sentinel-2, and LiDAR data," *Remote Sens.*, vol. 14, no. 2, p. 359, Jan. 2022.

[29] G. Shao, G. Shao, and S. Fei, "Delineation of individual deciduous trees in plantations with low-density LiDAR data," *Int. J. Remote Sens.*, vol. 40, no. 1, pp. 346–363, Jan. 2019.

[30] J. Picos, G. Bastos, D. Míguez, L. Alonso, and J. Armesto, "Individual tree detection in a eucalyptus plantation using unmanned aerial vehicle (UAV)-LiDAR," *Remote Sens.*, vol. 12, no. 5, p. 885, Mar. 2020.

[31] W. Chen, X. Hu, W. Chen, Y. Hong, and M. Yang, "Airborne LiDAR remote sensing for individual tree forest inventory using trunk detection-aided mean shift clustering techniques," *Remote Sens.*, vol. 10, no. 7, p. 1078, Jul. 2018.

[32] J. W. Atkins, A. E. L. Stovall, and C. A. Silva, "Open-source tools in R for forestry and forest ecology," *Forest Ecol. Manage.*, vol. 503, Jan. 2022, Art. no. 119813.

[33] W. Dai, B. Yang, Z. Dong, and A. Shaker, "A new method for 3D individual tree extraction using multispectral airborne LiDAR point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 400–411, Oct. 2018.

[34] L. Wallace, Q. Sun, B. Hally, S. Hillman, A. Both, J. Hurley, and D. S. M. Saldias, "Linking urban tree inventories to remote sensing data for individual tree mapping," *Urban Forestry Urban Greening*, vol. 61, Jun. 2021, Art. no. 127106.

[35] C. Zhang, Y. Zhou, and F. Qiu, "Individual tree segmentation from LiDAR point clouds for urban forest inventory," *Remote Sens.*, vol. 7, no. 6, pp. 7892–7913, Jun. 2015.

[36] J. Aval, J. Demuynck, E. Zenou, S. Fabre, D. Sheeren, M. Fauvel, K. Adeline, and X. Briottet, "Detection of individual trees in urban alignment from airborne data and contextual information: A marked point process approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 197–210, Dec. 2018.

LISHUO ZHANG received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in surveying and mapping science and technology from Tongji University, Shanghai, China, in 2019. He is currently working as a Postdoctoral Scholar with Sun Yat-sen University and the Guangzhou Urban Planning and Design Survey Institute. His research interests include forestry remote sensing and deep learning.

HONG LIN is currently the Vice President of the Guangzhou Urban Planning and Design Survey Research Institute. He is also a Professor in engineering investigation, the first batch of national registered surveyors engineering and the "special government allowances expert" in China. He has been worked on surveying for 34 years. His research interests include engineering survey and design. As the Project Leader, he won many significant science and technology awards, including three national excellent engineering survey and design industry awards.

FENG WANG received the Ph.D. degree in surveying and mapping science and technology from Tongji University, Shanghai, China, in 2009. He is currently the Minister of Research and Development Fellow with the Guangzhou Urban Planning and Design Survey Research Institute. He also has the title of a registered surveyor and a professional engineer. His research interests include point cloud processing and urban-scale 3D scene understanding.