TCDNet: Tree Crown Detection from UAV Optical Images Using Uncertainty-Aware One-Stage Network

Weichao Wu, Xijian Fan, *Member, IEEE*, Hongyu Qu, Xubing Yang, Tardi Tjahjadi, *Senior Member, IEEE*

***Abstract*—Tree crown detection plays a vital role in forestry management, resource statistics and yields forecasting. RGB high-resolution aerial images have emerged as a cost-effective source of data for tree crown detection. To address the challenges in the detection using UAV optical images, we propose a one-stage object detection network, TCDNet. First, the network provides an attention enhancement feature extraction module to enable the model to distinguish between tree crowns and their complex backgrounds. Second, an efficient loss is introduced to enable it to be aware of the overlap between adjacent trees, thus effectively avoiding misdetection. The experimental results on two publicly available datasets show that the proposed network outperforms state-of-art networks in terms of precision, recall and mean average precision.**

***Index Terms*—Convolutional neural networks, remote sensing, unmanned air vehicle, tree crown detection**

I. Introduction

Tree crown detection plays an important role in forestry and ecosystem research. The traditional methods of tree inventory involve field-based measurements, which is resource intensive and time demanding [1]. In recent years, remote sensing (RS) technology has been shown to be a less time-consuming tool for tree investigation. For this purpose, satellite, manned aircrafts and more recently, unmanned aerial vehicle (UAV) have been the most common platforms used for data acquisition. Satellite imagery (e.g., Sentinel-1, Sentinel2 and Landsat 8) suffers from insufficient spatial resolution, which renders the individual tree barely detectable [2, 3]. Manned aircraft is expensive and requires a huge amount of official paperwork and authorized licenses [4]. An UAV, e.g., a drone is considered to be an alternative, as it is cost-saving, lightweight, flexible, and easily manipulated.

Multi and hyperspectral images [5, 6], Light Detection and Ranging (LiDAR) data [7], or their combinations [8, 9] have been the preferred data source, and thus extensively investigated. However, these data are costly and could not be easily processed due to their high dimensionality. Alternatively, RGB images are cost efficient and easier to process in the absence of three-dimensional information of the tree crown.

Unfortunately, there are few studies on overlap in crown detection. All these studies directly exploit general object detection networks [10-12] based on deep convolutional neural network (CNN) [13] to detect tree crowns in RS scenes. Santos et al. [14] evaluated the performance of three detection methods, i.e., Faster R-CNN, YOLOv3 and RetinaNet, for individual tree detection. Weinstein et al. [15] also utilized RetinaNet to detect the tree crowns, achieving a promising precision. Oh et al. [16] applied a YOLO object detection model to detect and count cotton plants. Jintasuttisak et al. [17] exploited YOLO-V5 detection framework for detecting date palm trees. However, these methods fail to consider the issues that exclusively exist in the task of UAV based tree crowns detection as follows:

1) There exists the ambiguity in tree crowns detection using RS images, where it could be difficult to distinguish the trees of interest from their background as they tend to share the similar colour (e.g., grey-green) or under shade, as shown in Fig. 1(a).

2) Trees tend to grow densely, and trees that are close together may overlap. Furthermore, certain tree crowns are occluded due to camera angle view or distance, bringing more challenge to accurately detect individual crowns, as shown in Fig. 1(b).

(a) (b)

Fig. 1. Uncertainty in tree crown detection in RS scenes. Key: Green boxes denote the correct detection, and red boxes denote the missed and incorrect detection.

We refer the above two issues as the uncertainty in the task of tree crown detection with UAV, which requires to be considered when applying general object detectors to RS scenes. In the letter, we address such uncertainty by proposing an uncertainty-aware one-stage detection framework based on YOLO-X in [18]. We select YOLO-X based on its three prominent advantages: 1) Sensitive in detecting small objects (tree crowns in our task), avoiding the misdetection of some crowns with small size; 2) Robust to the variation of crown shape, which can generalize to different species of trees; and 3) Compared with two-stage detectors, it provides favourable tradeoff between accuracy and inference speed, which is suitable for UAV monitoring. To deal with the ambiguity that trees and background are not easy to distinguish, we make use of an attention strategy by integrating them to each branch at the end of the feature fusion module in YOLOX, enabling the model to pay more attention to the required tree crowns while

This work is supported by the National Natural Science Foundation of China under Grant 61902187, and in part by the Natural Science Foundation of Liaoning Province under grant 2020-KF-22-04 (*Corresponding author: Xijian Fan*)

Weichao Wu, Xijian Fan, Hongyu Xu, Xubing Yang are with the College of Information Science and Technology, Nanjing Forestry University, 210037 Nanjing, China (Wuweichao@njfu.edu.cn, xijian.fan@njfu.edu.cn, quhongyu@njfu.edu.cn, xbyang@njfu.edu.cn)

Tardi Tjahjadi is with the School of Engineering, University of Warwick, CV4 7AL, Coventry, UK (T.Tjahjadi@warwick.ac.uk)

alleviating the influence of complex background. In addition, we introduced an effective loss, which measures the extent of overlapping and occlusion between two crowns during the model learning, thus increasing the robustness of our model.

The main contributions of our work are summarized as follows:

1) We propose a robust and efficient model, TCDNet, specifically for tree crowns detection using UAV optical imagery. To the best of our knowledge, we are the first to address the uncertainty exclusively exists in RS tree crown detection.

2) We propose an attention enhancement PAN module, AEPAN, to highlight the significant feature information of the target tree crowns. This module further enhances the feature aggregation through fusion channel and spatial attention, thus focusing more on tree crowns while mitigating the impact of similar background and shade patterns.

3) We introduce an occlusion aware loss, which effectively reduces incorrect detection caused by the mutual overlap and occlusion between tree crowns.

4) To evaluate the superiority of the proposed network, we performed extensive experiments on two datasets containing tree crowns with varying sizes and from different geographical locations, comparing our method with other state-of-the-art methods qualitatively and quantitatively.

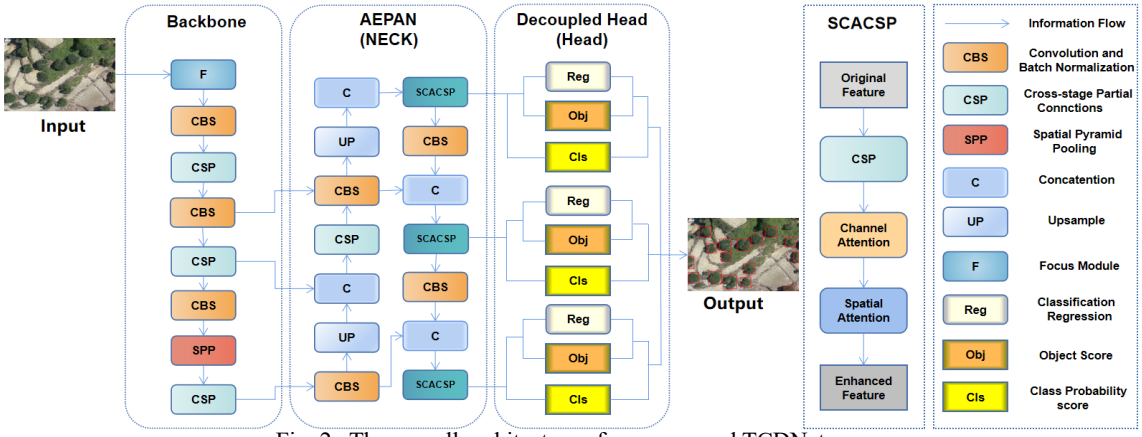

Fig. 2. The overall architecture of our proposed TCDNet.

II. Proposed Target Detection Method

The proposed TCDNet aims at solving the problems caused by the unique characteristics of RS tree crown images, e.g., the small size of tree crowns, the ambiguity regions of overlapping trees, and the similarity between crowns and background (or shade pattern) in RS images.

The overall framework of TCDNet (as shown in Fig. 2) comprises three main modules: Backbone using DarkNet53 with a spatial pyramid pooling (SPP) layer; AEPAN; and decoupled head. The DarkNet extracts the multi-scale features using SPP, which is able to capture the multi-scale representation of tree crowns from UAV images. The output multi-scale feature maps are then fed to the AEPAN module. AEPAN explores strong contextual features of the UAV images from top down and strong localization features of the target from the bottom up. These aggregate parameters of different detection layers to fully extract features. Finally, decouple head is used to identify the target locations and target class from the input features, where the refined loss function which takes into account the overlapping of tree crowns is designed to supervise the training process, thus increasing the robustness of the model.

A. Backbone

YOLOX is a one-stage object detection network newly proposed by the Megvii Research Institute [18], which achieves a promising performance in terms of inference speed and accuracy. We employ the YOLOX as our basic architecture. As an anchor-free detection method, there is no need to set anchor parameters. The backbone of our network is CSPDarkNet53, a combination of CSPNet and DarkNet53. It starts with the Focus module (F) (see Fig. 2) by transferring the spatial to the channel dimension without information loss. The CSP block consisting of a bottleneck structure and three convolutions is then used for feature extraction and selection. The feature maps from previous layer are computed into two parallel branches, and the number of channels is reduced to generate two new feature maps. The two maps are then concatenated as the output. The DarkNet53 architecture consists of 53 layers and contains 5 CSP blocks followed by convolution. It assumes the input image size is 640×640, and the size of successive output features are 320×320, 160×160, 80×80, 40×40 and 20×20, respectively. Moreover, a SPP layer is embedded to extract features at variable scales through maximum polling of different pooling kernel sizes. SPP improves the receptive field of the network and is also robust to the object deformation. In our work, we extract three feature maps of 80×80, 40×40 and 20×20 for feeding to the feature fusion.

B. Attention enhancement feature fusion module

Due to the characteristics of CNN that the more convolutional layers are stacked, the larger the receptive field of features is, and the more semantics and less low-level information are extracted. The classification of tree crowns heavily depends on semantics, and the localization is more

relevant to the receptive field of feature maps. To balance the classification and localization, a top-down and bottom-up multi-scale feature aggregation strategy PAN is exploited to fuse the features with different scales. However, due to the high similarity between tree crowns and their background in the RS images, the traditional PAN sometimes fails to capture the significant information of tree crowns, leading to a high probability of the background being detected as a target.

To solve the problem, we designed an attention enhancement PAN module, namely AEPAN. As shown in Fig. 3, the core module in AEPAN is a hybrid block that integrates an efficient attention mechanism to the CSP module of different branches, referred to as EACSP. The EACSP block allows the model to focus on the saliency of the tree crowns while suppressing background inference. We employ a dual attention strategy CBAM [19] due to its simplicity and effectiveness. This attention module is composed of two parts: spatial and channel-wise attention. Given an intermediate feature map in the CNN, CBAM injects the attention map along two independent dimensions, i.e., channel and spatial, and multiply the attention by input feature map to perform adaptive feature enhancement. Such attention enhancement feature is capable of capturing the important features and suppressing unnecessary ones. Fig. 2 illustrates the structure of SCACSP, where *H, W,* and *C* represent the dimensions of the feature map.

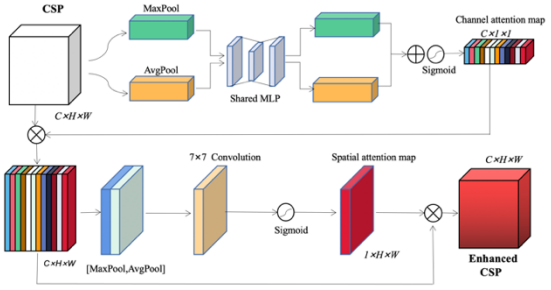

Fig. 3. Structure of SCACSP.

Specifically, the output features F of CSP module are first aggregated using global maximum pooling and average pooling to generate spatial descriptors F^C_{max} and F^C_{avg}. The two descriptors are then fed to a shared multi-layer perceptron network and merged using element-wise summation. The channel attention is computed as

$$\begin{aligned} \mathrm{CA(F)} &= \sigma(\mathrm{MLP(AvgPool(F))} + \mathrm{MLP(MaxPool(F))}) \\ &= \sigma(W_1(W_0(F^C_{avg})) + W_1(W_0(F^C_{max}))), \end{aligned} \quad (1)$$

where σ denotes the sigmoid function. W_0 and W_1 are the multilayer perceptron (MLP) weights that are shared for both inputs, where the ReLU activation function is followed by W_0. Multilayer Perceptron (MLP) represents two fully connected shared layers. The generated channel attention is multiplied by F to obtain the channel attention feature map F', i.e.,

$$F' = CA(F) \otimes F. \quad (2)$$

After obtaining F', the spatial attention is computed by applying average-pooling and max-pooling along the channel dimension, resulting in two descriptors F^S_{max} and F^S_{avg}. Such pooling operations highlight informative tree crown regions. The outputs are then concatenated to generate a feature descriptor and fed to a convolution layer to generate spatial attention map. The spatial attention is computed as

$$SA(F') = \sigma(Conv(F^S_{max}; F^S_{avg})). \quad (3)$$

The spatial attention map and F' are element-wise multiplied to obtain the final attention enhanced feature F_{AE}, i.e.,

$$F_{SCA} = SA(F') \otimes F'. \quad (4)$$

We integrate the attention strategy to all three CSP modules (with different feature scales), where multiscale salient feature information in tree crown RS images is well retained.

C. Decouple head and occlusion aware loss

Unlike other one-stage object detection methods, YOLOX replaces coupled detection heads with decoupled detection heads, which greatly improves the speed of convergence. In addition, dynamic top-k strategy named SimOTA and anchor-free detectors are also added [20]. After the feature fusion module, i.e., AEPAN, we forward three attention enhanced features to the decouple heads to obtain detection results.

The loss function of TCDNet contains three parts, i.e.,

$$L = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{bbox}, \quad (5)$$

where L_{cls} denotes the classification loss, L_{obj} denotes the objectness loss (reflecting the confidence score of object presence), and L_{bbox} denotes the bounding box regression score loss; and where λ_1, λ_2 and λ_3 are the weight coefficients (we set λ_1, λ_2 and λ_3 to 1, 1 and 5, respectively). The objectness loss is due to an incorrect prediction of box-object IoU, and teaches the network to predict a correct IoU, i.e., eventually pushing the IoU toward 1. In our case, the objectness and classification losses are calculated using binary cross-entropy loss (BCE), i.e.,

$$L(y_i, \hat{y}_i) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (6)$$

where y_i and $\hat{y}_i$ denotes the ground truth and prediction.

As shown in Fig.4, if we use IoU loss to calculate the regression loss of the bounding box, the IoU value between P and B is larger than the one between P and T. The box P will move to box B during the training, which causes the missed detection of the true box T. However, if the aspect ratio of the boundary is to be preserved, the box P is obviously more similar to box T. Thus, the estimated CIoU value between P and T is larger than the one between P and B, which moves the box P close to T during the training. The CIoU loss function L_{CIoU} is defined as

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v, \quad (7)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \quad (8)$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2, \quad (9)$$

$$\alpha = \frac{v}{(1 - IoU) + v}. \quad (10)$$

where b and b^{gt} respectively denote the centre points of the predicted and the ground-truth box. ρ is the Euclidean distance between the two centre points, and c denotes the diagonal distance of the smallest closure area that contains both the predicted box and the ground-truth box. α is the weight function and the penalty term v is used to measure the similarity via the aspect ratio. w and h are respectively the width and height of bound box. CIoU loss takes into account the distance between the target and the anchor, the overlap rate and the scale, which thus makes the target box regression more stable and robust to overlapping and occlusions.

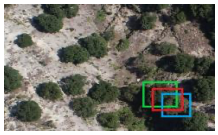

Fig.4. Illustration of the CIoU Loss. Key: Green box denotes ground truth T, blue box denotes the overlapping ground truth B, and red box denotes prediction P.

III. Experiments and Results

A. Datasets

We used two publicly available datasets, Neon Tree Crowns dataset [21] and Bayberry Tree dataset [22]. For Neon Tree Crowns dataset, we *downloaded and collated 169 images of size 400×400 with the corresponding annotations from https://zenodo.org/record/59145 54#.YwYeJXZBy3C,which is a subset of Neon Tree Crowns dataset. This dataset contains canopy images of different species in multiple regions. For each geographic site a NEON four letter code (e.g HARV -> Harvard Forest) is provided. The Bayberry Tree dataset consists of 284 high-definition RS images with labels and size of 1024×682, acquired during January 23rd to 24th, 2019 from Dayangshan Forest Park, Yongjia County, Zhejiang Province. The images are collected using DJI Phantom4 drone for aerial photography. In each dataset, we confirm the accuracy of the annotation of each canopy, and each image is provided with geographic information about the tree in the image.*

B. Training Details

We conducted all experiments using Pytorch toolbox on a PC with NVIDIA Tesla P100. We randomly divided images from two datasets into training and test set at a ratio of 7:3, respectively. The size of input images is scaled to 640 × 640, using RandomAffine, RandomFlip, and the contrast conversion online enhancement strategy for data augmentation. The model loads the weights trained on the VOC dataset during model training. To protect the weights from being destroyed, we trained the network for 50 epochs by freezing feature extraction layers with an initial learning rate of 0.001 and a batch size of 4. We then trained the entire network with a learning rate of 0.0001, a mini-batch size of 4 for 50 epochs. During the training, we use the Adam optimizer, and set the momentum and weight decay to 0.9 and 0.0005, respectively. In all experiments, we set the IOU threshold to 0.5. When the overlapping area between the prediction box and the ground-truth exceeds 50%, the prediction box is considered to be correct. When calculating all metrics results, we set the confidence threshold to 0.5.

C. Evaluating Metric

To quantitatively assess the performance of our TCDNet for detecting tree crowns, the general three metrics of recall, precision and mean average precision (mAP) are used. Recall is defined as the ratio of correctly detected objects by the model in all ground-truth objects, and precision refers to the ratio of the correctly detected objects by the model in all detected objects. The recall and precision rates are computed as

$$Precision = \frac{TP}{TP+FP}, \tag{11}$$

$$Recall = \frac{TP}{TP+FN}, \tag{12}$$

where *TP* denotes the number of tree crowns correctly detected, *FP* denotes the number of other objects detected as tree crowns, and *FN* is the number of tree crowns that are not detected.

The mAP represents the average value of each class average precision (AP), i.e., the area under the precision-recall rate (R-R) curve, which is defined as

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 P(R)dR, \tag{13}$$

where *N* denotes the number of detected classes. In our case, as *N* is 1, mAP is equivalent to AP.

D. Experimental Results and Analysis

In order to evaluate the effectiveness of the proposed method, different state-of-the-art object detection networks are employed for comparison, including Faster-RCNN [11], SSD [27], YOLOV3, YOLOV4 and YOLOX. Each method was trained with two datasets using the VOC pre-trained weights. The experimental results are shown in Table 1.

TABLE I. PERFORMANCE COMPARISONS ON TWO DATASETS

Datasets	Model	Precision	recall	mAP
Neon Tree Crowns	Faster-RCNN	50.17%	31.57%	39.73%
	SSD	63.47%	33.72%	45.83%
	YOLOV3	67.63%	38.11%	49.14%
	YOLOV4	65.33%	43.03%	54.34%
	YOLOX	69.57%	45.25%	57.99%
	TCDNet	**70.53%**	**46.90%**	**58.49%**
Bayberry Tree	Faster-RCNN	77.62%	70.55%	75.12%
	SSD	85.17%	80.76%	82.45%
	YOLOV3	80.15%	86.61%	87.31%
	YOLOV4	86.21%	92.68%	93.52%
	YOLOX	87.34%	94.80%	96.30%
	TCDNet	**87.98%**	**95.41%**	**96.76%**

There are two key modules within TCDNet, i.e., AEPAN module and CIoU loss. To validate the effectiveness of these two improvements, we perform an ablation study. In the proposed model, AEPAN aggregates the salient information by adding the attention module. CIoU loss is introduced to replace IoU loss, enabling the network to be aware of the relation between close trees. We performed 4 sets of ablation experiments: 1) Using baseline YOLOX; 2) Replacing PAN with AEPAN; 3) Replacing PAN with AEPAN and using CIoU loss; and 4) Using TCDNet. All hyperparameters were retained constant throughout the experiment.

TABLE II. RESULTS OF ABLATION STUDY

Datasets	Model	Precision	Recall	mAP
Neon Tree Crowns	YOLOX	69.57%	45.25%	57.99%
	AEPAN	70.24%	45.67%	58.39%
	CIoU	70.12%	45.49%	58.32%
	TCDNet	**70.53%**	**45.90%**	**58.49%**
Bayberry Tree	YOLOX	87.34%	94.80%	96.30%
	AEPAN	87.11%	95.07%	96.42%
	CIoU	87.51%	95.23%	96.62%
	TCDNet	**87.98%**	**95.41%**	**96.76%**

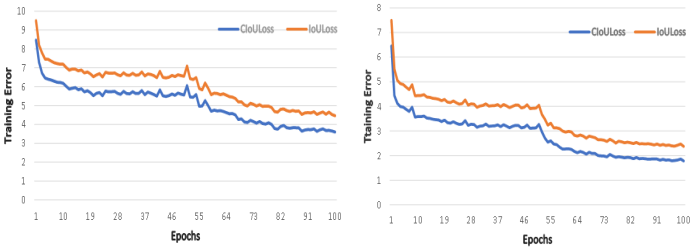

Fig. 5. The loss curve during training (Left): Training on Neon Tree Crowns dataset; and (Right): Training on Bayberry Tree dataset.

Table II shows that the proposed TCDNet achieves a constantly improvement in terms of precision, recall and mAP. Compared with YOLOX using PAN, the introduction of AEPAN module increases the mAP by 0.40% and 0.12% on the two datasets, respectively. The CIoU loss function improves mAP on the two datasets by 0.33% and 0.32%, respectively.

Furthermore, the value of CIoU loss is constantly lower than IoU on both datasets during training (shown in Fig. 5). Overall, the proposed TCDNet using AEPAN and CIoU has greatly improved the performance for tree crowns detection using RS imagery compared with the original YOLOX.

In order to show the effectiveness of our methods compared with the original YOLOX, we visualized the detection results in the images under different environments using YOLOX and our method, respectively. The experimental results are shown in Fig. 6. The first row and the second row are the detection results using YOLOX and TCDNet, receptively. There are serious problems such as miss detection and inaccurate detection using YOLOX. This is due to: 1) High similarity between tree crowns and their background; and 2) Overlapping of tree crowns. Our TCDNet has succeeded in solving these two problems and detects almost all the targets missed by YOLOX in the images from both two datasets. Therefore, the TCDNet has excellent detection performance in terms of accuracy and robustness, which is well generalized to other RS scenarios

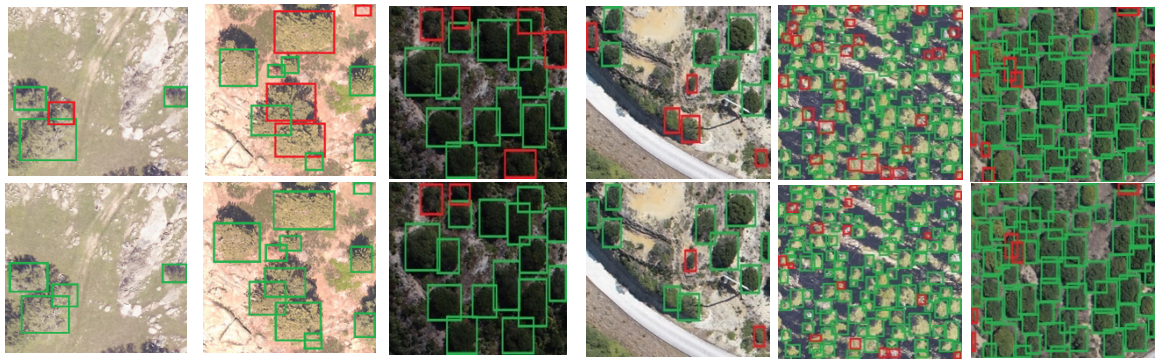

Fig. 6. Comparison of detection results of using YOLOX and TCDNet on two datasets. Images of first two columns from Neon Tree Crowns dataset; and images from the remaining columns from Bayberry dataset. The top row shows detection results using the original YOLOX, and the bottom row shows detection results using our TCDNet. Key: Green boxes denote the correct detection, and red boxes denote the missed and incorrect detection.

IV. Conclusion

We proposed a one-stage object detection network to detect tree crown using UAV imagery. Aiming at the problems of missed/false detection caused by the high similarity between tree crowns and their background, the attention enhanced module is integrated to the multi-scale feature fusion layer. Using the CIoU loss instead of the IoU loss enables the model more robust to the missed detection due to the occlusion between tree crowns. The experimental results on two datasets verify the effectiveness of each module was designed as well as the entire framework. In future, we will apply our method to the practical applications of forestry resource surveys.

REFERENCES

[1] M. Campbell, P. Dennison, J. Tune, et al., "A multi-sensor, multi-scale approach to mapping tree mortality in woodland ecosystems," *Remote Sens Environ*, vol. 245, 111853, 2020.

[2] C. R Axelsson & N. P. Hanan, Biogeosciences vol. 14, pp.3239–3252, 2017.

[3] A. Matese et al., "Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture," Remote Sens., vol. 7, no. 3, pp. 2971–2990, Mar. 2015.

[4] M Miraki, H Sohrabi, P Fatehi, et al. "Individual tree crown delineation from high-resolution UAV images in broadleaf forest," Ecological Informatics, vol. 61, 101207, 2021

[5] J. Maschler, C. Atzberger, M. Immitzer, "Individual tree crown segmentation and classification of 13 tree species using airborne hyperspectral data," *Remote Sens.*, vol. 10, no. 8, pp. 1218, 2018.

[6] R. Minařík, J. Langhammer, T. Lendzioch, "Automatic tree crown extraction from UAS multispectral imagery for the detection of bark beetle disturbance in mixed forests," *Remote Sens*., vol. 12, no. 24, pp. 4081, 2020.

[7] D. Marinelli, C. Paris and L. Bruzzone, "An Approach to Tree Detection Based on the Fusion of Multitemporal LiDAR Data," *IEEE Geosci. Remote Sens. Lett*., vol. 16, no. 11, pp. 1771-1775, Nov. 2019.

[8] M. Dalponte, L. Bruzzone, D. Gianelle. "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data," *Remote Sens. Environ.*, vol. 123, pp. 258-270, 2012.

[9] D. Marinelli, C. Paris, L. Bruzzone, "An approach to tree detection based on the fusion of multitemporal LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1771-1775, 2019.

[10] T Lin, P. Goyal, R, Girshick, et al., "Focal loss for dense object detection," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2980-2988, Octorber, 2017.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, pp. 2891–2899, December, 2015.

[12] J. Redmon, S. Divvala, R. Girshick, et al., "You only look once: Unified, real-time object detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 779-788, 2016.

[13] A. Krizhevsky, I. Sutskever, G. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 25, December, 2012.

[14] A. Santos, J. Junior, M. Araújo, et al., "Assessment of CNN-Based Methods for Individual Tree Detection on Images Captured by RGB Cameras Attached to UAVs," *Sensors*, vol. 19, no. 16, pp. 3595, 2019.

[15] B G Weinstein, Marconi S, Bohlman S, et al. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks[J]. Remote Sensing, 2019, 11(11): 1309.

[16] S. Oh, A. Chang, A. Ashapure, et. al., "Plant Counting of Cotton from UAS Imagery Using Deep Learning-Based Object Detection Framework," *Remote Sens.*, vol. 12, no. 28., pp. 2981, 2020.

[17] T. Jintasuttisak, E. Edirisinghe, and A. Elbattay, "Deep neural network based date palm tree detection in drone imagery," *Comput. Electron. Agric.*, vol. 192, 106560, 2022.

[18] Z. Ge, S. Liu, F. Wang, et. al., "Yolox: Exceeding yolo series in 2021," arXiv 2021, arXiv:2107.08430, 2021.

[19] S. Woo, J. Park, J. Lee, et al., "Cbam: Convolutional block attention module" *Proc. Europ. Conf. Comp. (ECCV)*, pp. 3-19, 2018.

[20] Zheng Z, Wang P, Ren D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021.

[21] B. G. Weinstein, S. Marconi, S. Bohlman, et al., "NEON Crowns: a remote sensing derived dataset of 100 million individual tree crowns, bioRxiv 2020.

[22] D. Wang, W. Luo, "Bayberry tree recognition dataset based on the aerial photos and deep learning model," J Global Change Data Discover, vol. 3, no. 3, pp. 290-296, 2019.

[23] W. Liu, D. Anguelov, D. Erhan, et al., "Ssd: Single shot multibox detector," Proc. Europ. Conf. Comp. (ECCV), pp. 21-37, 2016.