Tree Species Identification and Counting in UAV Optical Images Based on Improved YOLOv8n

Chenyang HU, Jie XU*

Heilongjiang Bayi Agricultural University, Daqing 163000, China

Abstract Forests play a crucial role in ecosystems. This study focused on five common tree species in Northeast China: pine, elm, poplar, cedar, and ash. An improved YOLOv8n-based network structure was constructed, and a UAV image dataset was developed for analysis. The results showed that the improved YOLOv8 algorithm achieved a 4.9% increase in accuracy compared with the original version, and the average precision increased from 88.0% (original YOLOv8n) to 92.1%.

Key words YOLOv8n; UAV; Module; Tree species identification

DOI:10.19759/j.cnki.2164 – 4993.2025.04.018

Forests play a vital role in ecosystems[1]. They not only help maintain the water cycle, protect soil and absorb carbon emissions, but also provide habitats for humans and animals[2]. To address the issue of low accuracy in tree recognition and counting in complex backgrounds[3], this paper introduced two modules, namely, the CMEA (Convolutional Multi-scale Enhancement Attention) module and the DPSA (Dynamic Pooling Synergy Attention) module, to improve the precision of tree species recognition and counting by YOLOv8 in complex natural environments. The CMEA module aims to capture feature information at different scales by introducing multi-scale convolutional layers, while integrating channel attention and spatial attention mechanisms for adaptive feature enhancement, thereby improving the models' performance in handling various visual tasks. The DPSA module enhances the expressive power of input feature maps and improves adaptability to multi-scale features. Embedded in the backbone of YOLOv8n, they enable efficient device deployment and provide technical support for the digital management of forest resources.

Method and Algorithm Design

Improved YOLOv8 network model design

In this study, an enhanced YOLOv8-based network architecture tailored for tree species identification and quantity estimation tasks in high-resolution UAV images was proposed[4]. Compared with traditional remote sensing classification methods, YOLOv8 offers advantages such as end-to-end modeling, fast inference speed, and strong multi-object detection capability[5], making it suitable for real-time identification of tree species in densely forested areas. However, due to the high visual similarity between different tree species (*e.g.*, leaf texture, canopy shape) and challenges in UAV images such as significant scale variations, uneven lighting, and occlusion interference[6], the original YOLOv8 still faces performance bottlenecks in high-precision tree species recognition[7].

For this purpose, this study proposed the integration of two innovative modules, the CMEA (Convolutional Multi-scale Enhancement Attention) module and the DPSA (Dynamic Pooling Synergy Attention) module into the original structure of YOLOv8 (Fig. 1).

Convolutional multi-scale enhancement attention module (CMEA)

The CMEA (Convolutional Multi-scale Enhancement Attention) module is designed to address insufficient feature extraction caused by similar textures among tree species and large scale variations in UAV images. The module integrates multi-scale convolutional kernel structures with a dual attention mechanism (channel and spatial). Through multi-path parallel convolution processing on input feature maps, it achieves information extraction in different receptive fields, followed by reconstruction and enhancement of high-dimensional semantic features through fusion operations.

As shown in Fig. 2, the module first processes the input feature map using multiple groups of depthwise separable convolutions with different kernel sizes (3 × 3, 5 × 5, 7 × 7, 9 × 9) to extract multi-scale features. These features are then concated in the channel dimension and compressed/fused via 1 × 1 convolution. The fused features are subsequently fed into two submodules:

Channel attention branch: Channel descriptor vectors are generated through global AvgPool and MaxPool, and the channel weights are computed via a multilayer perceptron (MLP) to enhance key channel features. Spatial attention branch: The features fused in the channel dimension undergo AvgPool and MaxPool in the channel direction before being concated, and spatial dependency feature extraction is performed through a 7 × 7 convolution to generate a spatial attention map that emphasizes critical regions (*e.g.*, tree crown center and edge transition zones). Finally, the two attention maps are normalized using the Sigmoid function, multiplied with the backbone features, and then summed to output

Received: March 29, 2025 Accepted: May 31, 2025

Chenyang HU (2000 –), male, P. R. China, master, devoted to research about forestry information.

* Corresponding author.

the enhanced multi-scale representation. The overall enhancement process of the CMEA module can be simplified as the following equation:

$$X_{\text{final}} = X_{\text{fused}} \cdot S_{\text{attention}} + X_{\text{fused}} \cdot C_{\text{attention}}$$

In the equation, X_{final} represents the multi-scale convolution fusion features; and $S_{\text{attention}}$ and $C_{\text{attention}}$ denote, respectively, the spatial and channel attention weight maps, which jointly strengthen salient regions and critical feature channels.

The introduction of the CMEA module significantly enhances the model's perception capability for targets with large-scale variations (*e. g.*, saplings, occluded canopies), while improving focus on key regions through the joint channel-spatial attention mechanism. Experiments confirm its effectiveness in reducing false detections caused by background clutter or uneven illumination.

Dynamic pooling synergy attention module (DPSA)

To further enhance feature fusion and semantic representation in complex scenes, this paper introduced the DPSA (Dynamic Pooling Synergy Attention) module into the Neck stage of YOLOv8, aiming to strengthen the interactivity and discriminative capability among multi-scale feature maps.

As shown in Fig. 3, the DPSA module first extracts statistical features at different perceptual levels through three pooling operations (MaxPool, AvgPool, and MixPool). A shared MLP structure then generates weighted channel attention maps to achieve selective enhancement of feature channels. Compared with traditional attention mechanisms, the MixPool strategy adaptively adjusts the proportion of each pooling operation, enabling the model to respond flexibly when processing both large-canopy trees and small-sized forest trees.

Subsequently, the DPSA module incorporates an MSDC (Multi-Scale Dilated Convolution) branch, which employs four groups of parallel dilated convolution structures with dilation rates of 1, 3, 5 and 7 to extract long-range contextual information, enhancing global perception for large-scale targets. The outputs of the four dilated convolution groups are concated in the channel dimension, and then processed by a convolutional MLP to unify dimensions before being fed into the spatial attention branch.

The spatial attention adopts a structure similar to the CMEA module. After performing MaxPool and AvgPool in the channel dimension and following concatenation, a 7 × 7 convolution generates a spatial weight map to enhance responses in salient regions of the feature map. Finally, the channel attention and spatial attention results are multiplied and applied to the backbone features, producing the fusion expression. The DPSA module ultimately combines channel pooling attention, dilated convolution, and spatial attention for enhancement. Its overall output can be expressed as:

$$output = \sigma\ (Conv\ (conact\ ([Mean\ (I,\ dim = 1),\ Mean\ (I,\ dim = 1)]))).\ MSDC\ (z.\ \sigma\ (f_{avg} + f_{max} + f_{min}))$$

In the equation, I denotes the input features; f_{avg}, f_{amx} and f_{min} represent the channel attention outputs from the three pooling methods, respectively; z is the scaling factor; *MSDC* denotes the multi-scale dilated convolution; σ is the Sigmoid activation function; and the output is the final fused feature map.

The DPSA module employs a multidimensional fusion strategy of "channel-space-context" to effectively enhance the Neck stage's capacity to aggregate multi-scale semantic information. It is particularly suited for addressing challenges such as occlusion by large-canopy trees and weak responses to targets at image edges. This module improves the model's localization accuracy in high-density areas while suppressing background interference and reducing missing detection.

The channel attention mechanism aims to learn channel importance through pooling operations. The channel weights are generated by combining AvgPool, MaxPool, and MixPool.

Dataset Construction

The dataset used in this study was a self-collected dataset. The data were acquired in June 2024 from the forest farm of Northeast Forestry University in Harbin, Heilongjiang Province, China. As shown in Fig. 4, aerial images of the forest area were captured vertically using drones outdoors. The flight altitude was maintained at 80 m, and a high-resolution camera was employed to reduce the risk of image blur. The flight route was set with an 80% forward overlap rate and 75% side overlap rate to facilitate subsequent image stitching. Image acquisition was conducted daily between 14:00 and 17:00 when the light was soft and shadows were minimal, ensuring clear visualization of canopy texture features.

To ensure data diversity and representativeness, five typical tree species were selected as primary research subjects: pine, poplar, elm, cedar, and ash. The drones followed predefined flight paths to conduct comprehensive coverage shooting of multiple sample plots, while simultaneously collecting corresponding ground truth data (including tree species type, GPS coordinates, and quantity annotations) for subsequent data labeling and validation.

After image acquisition, an OpenCV-based frame extraction script was first developed to process the original aerial videos by extracting frames at 3 fps, yielding 4 283 initial images. To eliminate duplicates and invalid samples (*e. g.*, blurred images, occluded canopies, or repeated empty areas), the Structural Similarity Index (SSIM) algorithm was applied for similarity filtering with a threshold of 0. 85. This process removed 1 027 redundant images, leaving 3 256 valid samples.

Subsequently, the LabelImg annotation tool was employed to label each image individually, with target location information output in YOLO format. The annotations covered five tree species, totaling 10 244 labeled targets, averaging 3. 1 targets per image.

The finalized tree species recognition dataset consisted of 3 256 images, divided into training (2 604 images), validation (326 images), and test sets (326 images) at an 8 : 1 : 1 ratio for subsequent model training and performance evaluation. Fig. 5 displays representative sample images from the collected dataset.

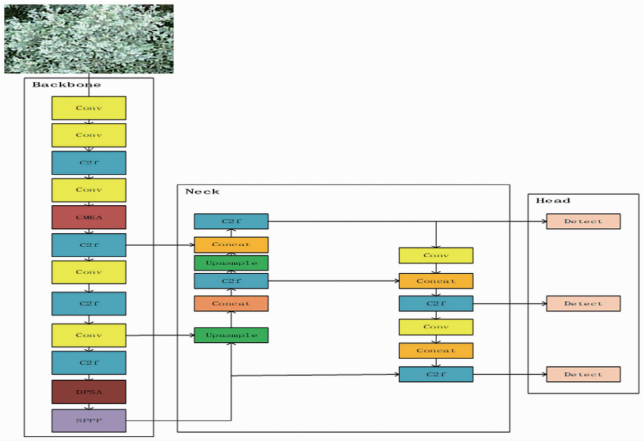


Fig. 1 Improved YOLOv8 network structure diagram

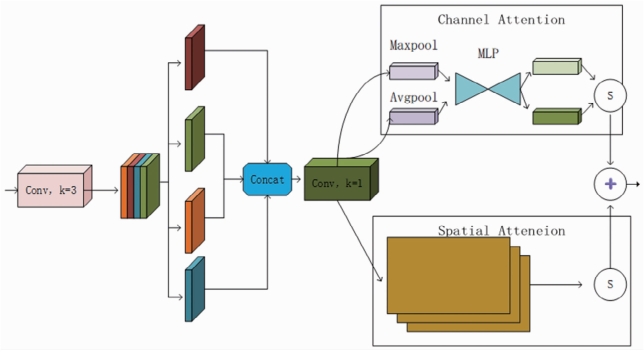


Fig. 2 CMEA module structure diagram

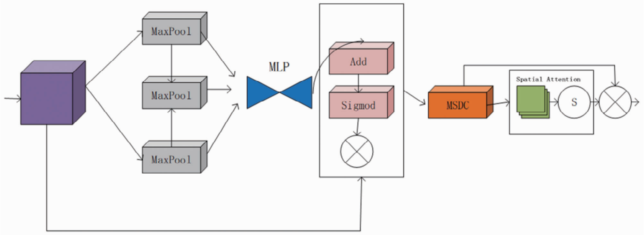
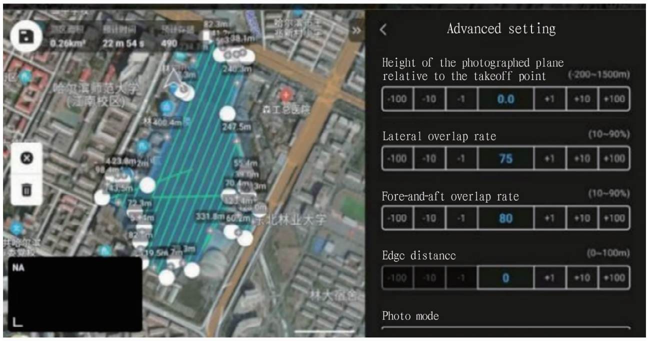


Fig. 3 DPSA module structure diagram



Advanced setting

Height of the photographed plane relative to the takeoff point (-200~1500m)
-100 -10 -1 **0.0** +1 +10 +100

Lateral overlap rate (10~90%)
-100 -10 -1 **75** +1 +10 +100

Fore-and-aft overlap rate (10~90%)
-100 -10 -1 **80** +1 +10 +100

Edgc distancc (0~100m)
-100 -10 -1 **0** +1 +10 +100

Photo mode

Fig. 4 Basic information about drones

Results and Analysis

Experimental environment

The experiments were conducted on a Windows 11 operating system, and the hardware platform was a portable high-performance computing device equipped with an NVIDIA GeForce RTX 4060 Laptop GPU. The model was trained using the Python 3.10 programming language and the PyTorch deep learning framework. The training was set to 200 epochs with an initial learning rate of 0.01, a batch size of 4, and an input image size of 640×640 pixels. The stochastic gradient descent (SGD) optimizer was employed. All training processes were conducted under identical data augmentation strategies and data splits to ensure the fairness of comparative experiments.

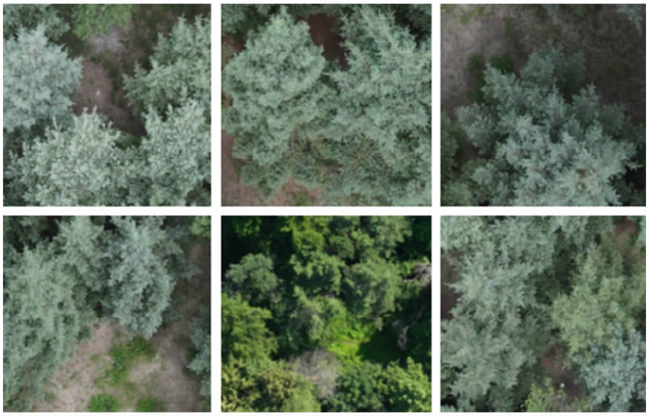


Fig. 5 Representative image example

Module improvement comparison experiments and results

To validate the improvement effects of the CMEA and DPSA modules in the YOLOv8 network, four modified models were constructed in this study: YOLOv8-CMEA, YOLOv8-DPSA, YOLOv8-CBAM-CMEA, and YOLOv8-CMEA-DPSA, which were compared with the original YOLOv8n. Table 1 presents the key performance indices of each model in the tree species recognition task, including Precision, mAP@0.5, Recall, and inference speed (FPS).

Table 1 Comparison of YOLOv8 model improvements

Model name	Precision	mAP@0.5	Recall	FPS
YOLOv8n (original)	0.885	0.880	0.868	63.4
YOLOv8n-CMEA	0.895	0.884	0.873	65.2
YOLOv8n-DPSA	0.901	0.885	0.873	66.3
YOLOv8n-CBAM-CMEA	0.905	0.885	0.893	68.3
YOLOv8n-CMEA-DPSA	0.934	0.921	0.882	70.1

In this study, three optimization algorithms were proposed: CMEA, DPSA, and CBAM. Compared with several attention mechanism module modifications, the model integrating both CMEA and DPSA attention mechanisms performed particularly well. The improved model significantly outperformed the original YOLOv8n model in terms of precision, mean average precision (mAP), recall, and frame rate. Experimental results showed that the precision of the model increased from 0.885 to 0.934, the mAP increased from 0.880 to 0.921, and the recall increased from 0.868 to 0.882.

To better demonstrate the effectiveness of our proposed model, the object detection results was presented in this study. The improved model maintained robust recognition performance even in complex environments (Fig. 6).

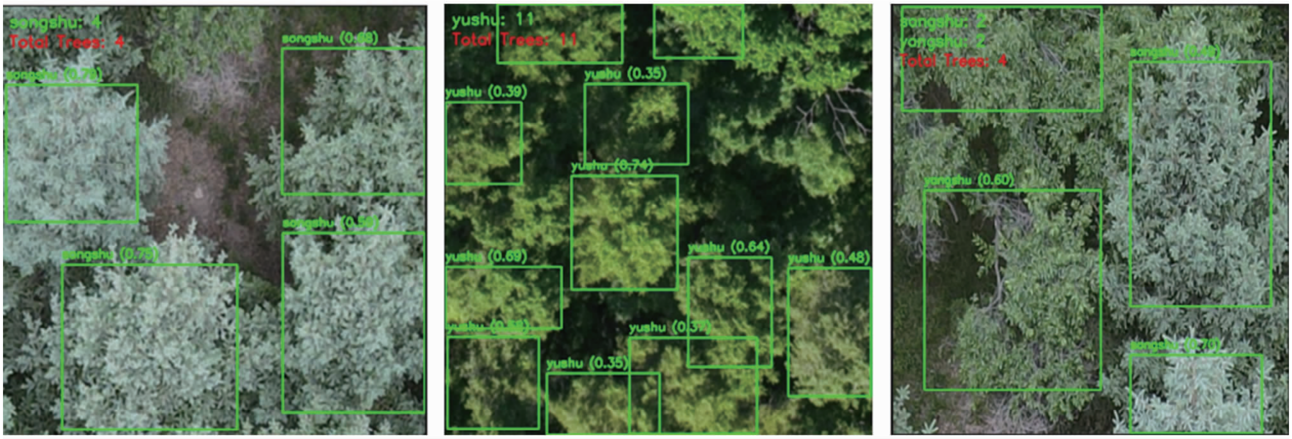


Fig. 6 Recognition performance of the YOLOv8n-CMEA-DPSA model

Conclusions and Discussion

In this study, the issue of low accuracy in tree species identification and counting in UAV optical images under complex backgrounds was addressed by proposing an improved model based on YOLOv8, incorporating two attention modules, CMEA and DPSA. The experimental results demonstrated that the improved YOLOv8n-CMEA-DPSA model significantly outperformed the original model in terms of accuracy, average precision, and recall, and the average precision increased to 92.1%, fully validating the practical application potential of the proposed method in forest resource monitoring. This study provides technical support for digital forest management and explores new approaches for applying deep learning to high-resolution remote sensing images.

References

[1] XIE YN. Research on tree species identification of UAV multi-source data based on deep learning[D]. Nanjing: Nanjing Forestry University, 2023. (in Chinese).

[2] GAO X. Research on tree species identification based on UAV multi-source data and deep learning[D]. Harbin: Northeast Forestry University, 2020. (in Chinese).

[3] DING Y, XU AJ, WU XF, *et al.* Common arbor identification method in subtropics based on multiple features fusion and knowledge distillation [J]. Application of Electronic Technique, 2024, 50(8): 1 – 9. (in Chinese).

[4] ZHENG ZL. Urban tree species identification and biomass estimation based on UAV multi-source remote sensing data[D]. Kaifeng: Henan University, 2024. (in Chinese).

[5] SHI H, WANG Y, FENG X, *et al.* YOLOv8-MFD: An enhanced detection model for pine wilt diseased trees using uav imagery[J]. Sensors, 2025, 25(11): 3315.

[6] YANG T, ZHOU S, XU A, *et al.* YOLO-SegNet: A method for individual street tree segmentation based on the improved YOLOv8 and the SegFormer network[J]. Agriculture, 2024, 14(9): 1620.

[7] ZHANG L, YU S, YANG B, *et al.* YOLOv8 forestry pest recognition based on improved re-parametric convolution[J]. Frontiers in Plant Science, 2025, 16: 1552853.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU

(Continued from page 83)

based on learning big data)[4–5]. We will deepen industry-academia collaboration to develop more cutting-edge virtual simulation and case resources[6], so as to establish a new digital teaching ecosystem. These efforts will sustainably enhance the cultivation of high-quality applied bioengineering talents meeting engineering certification standards to provide solid talent support for the development of Southwest China's bioindustry.

References

[1] WANG B, ZHANG T, LI M, *et al.* Teaching reform of molecular biology theory course and experiment course based on OBE educational concept [J]. The modern occupation education, 2021(33): 12 – 13. (in Chinese).

[2] QIN XH, CHEN ZW, WANG X. Research and practice of "molecular biology experiment" course based on OBE teaching concept and PBL teaching mode[J]. Technology Wind, 2023(17): 103 – 105. (in Chinese).

[3] HE SJ, XIONG YJ, LI J, *et al.* Research on the evaluation of teachers' teaching quality under the OBE education concept[J]. Contemporary Animal Husbandry, 2022(4): 78 – 80. (in Chinese).

[4] LI ZH, WANG HF, LIU CY, *et al.* Application and practice of artificial intelligence empowering molecular biology teaching[J]. Chinese Journal of Biochemistry and Molecular Biology, 2025: 1 – 14. (in Chinese).

[5] LI MY, WU LT, LAN X, *et al.* Application prospect of generative artificial intelligence in the experimental teaching of biochemistry[J]. Basic Medical Education, 2025, 27(4): 329 – 332. (in Chinese).

[6] SHAO H, CHI L, HU XM, *et al.* Exploration of teaching reform of biopharmaceutical course under the background of "curriculum ideology and politics" and "internet +"[J]. Education Forum, 2024(31): 97 – 100. (in Chinese).

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU