GACNet: A Geometric and Attribute Co-Evolutionary Network for Citrus Tree Height Extraction From UAV Photogrammetry-Derived Data

Haiqing He, Fuyang Zhou, Yongjun Zhang, *Member, IEEE*, Ting Chen, and Yan Wei

***Abstract*—The undulating terrain and complex backgrounds of citrus plantations introduce nonlinear variations that significantly impede the high-precision estimation of citrus tree heights from remote sensing data. To overcome these obstacles, we introduce a novel geometric and attribute co-evolutionary network, tailored for extracting citrus tree heights using unmanned aerial vehicle photogrammetry-derived data. Our approach integrates a multisource feature interaction module with a multisource feature aggregation module, fostering the co-evolution of deep feature responses across various datasets. Notably, this includes a sophisticated triple-feature interaction mechanism that considers position, channel, and spatial correlation to enhance the aggregation of geometric features. In addition, we employ a multilevel feature aggregation decoder leveraging cross-attention, ensuring attribute context consistency and facilitating efficient tree height extraction. Quantitative analysis across datasets reveals our method's superior performance, with a 2% –7% increase in mean intersection over union for canopy segmentation and a robust correlation of 0.77 between estimated and reference tree heights, accompanied by an MAE of 0.25 m and an RMSE of 0.38 m. Comparative experiments indicate that our method outperforms current state-of-the-art networks, showing resilience to terrain undulations and offering reliable cross-region and cross-scale tree height estimation capabilities.**

***Index Terms*—Canopy height, co-evolutionary network, feature interaction and aggregation, undulating terrain, unmanned aerial vehicle (UAV).**

Received 13 September 2024; revised 19 November 2024; accepted 8 February 2025. Date of publication 13 February 2025; date of current version 3 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42261075 and Grant 41861062, in part by the Jiangxi Provincial Natural Science Foundation under Grant 20224ACB212003, in part by the Jiangxi Provincial Training Project of Disciplinary, Academic, and Technical Leader under Grant 20232BCJ22002, and in part by the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, Chinese Academy of Surveying and Mapping under Grant 2022-02-04. *(Corresponding author: Haiqing He.)*

Haiqing He, Fuyang Zhou, and Yan Wei are with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China, and also with the Jiangxi Key Laboratory of Watershed Ecological Process and Information (Platform No. 2023SSY01051), East China University of Technology, Nanchang 330013, China (e-mail: hyhqing@163.com; fuy_zhou@163.com; ywei0623@163.com).

Yongjun Zhang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zhangyj@whu.edu.cn).

Ting Chen is with the School of Water Resources and Environmental Engineering, East China University of Technology, Nanchang 330013, China (e-mail: ct_201607@ecut.edu.cn).

Data is available online at https://doi.org/10.1109/JSTARS.2025.3541395.

Digital Object Identifier 10.1109/JSTARS.2025.3541395

I. Introduction

CITRUS is one of the most important economic crops globally, ranking as the largest category of fruits and the third-largest traded agricultural commodity worldwide. Currently, citrus cultivation is primarily concentrated in Asia, with its cultivation area accounting for 52.90% of the total global citrus cultivation area [1]. Particularly, citrus trees represent a representative fruit tree in southern China, exerting significant economic and ecological impacts, and have emerged as one of the vital sources of income for local residents in southern China [2], [3]. Accurate and rapid acquisition of canopy area, height, and positional information of citrus trees is crucial for monitoring tree health, estimating citrus yield, and managing orchard resources effectively [4]. However, existing technological methods face challenges in automatically and efficiently acquiring growth status information of citrus orchards, thereby hindering the implementation of precision management in citrus cultivation. Given that the terrain in southern China is predominantly hilly and mountainous, characterized by significant fluctuations and dense tree growth, monitoring large areas of citrus orchards dynamically poses difficulties [5]. Traditional field survey methods can offer reliable canopy information of citrus trees but are constrained by limited measurement range, long intervals, and high costs, thus failing to meet the requirements of precise monitoring of citrus orchards [6]. Therefore, in this study, we focus on achieving high-precision and efficient extraction of citrus trees that are insensitive to terrain undulations, aiming to enhance the accuracy of canopy and height extraction for large-scale citrus trees in complex terrains.

In terms of data acquisition techniques, light detection and ranging (LiDAR) is one of the most prevalent methodologies for the measurement of canopy and height within extensive forested regions [7]. The three-dimensional point cloud data derived from LiDAR apparatuses can precisely estimate numerous biophysical variables within forests, including canopy area, canopy height, canopy volume, and tree density [8], [9], [10], [11], [12]. However, there are certain limitations associated with the estimation of tree canopy height using LiDAR data, such as high costs, limited coverage, and inability to acquire color and texture information, which hampers the interpretation of terrestrial objects, leading to challenges in widespread application within complex mountainous terrain scenarios [13], [14], [15], [16]. To address these issues, some studies have integrated

satellite imagery with LiDAR data to estimate the area and height of large-scale tree canopies. These methodologies offer advantages such as extensive coverage and reduced data acquisition costs. Nevertheless, there are constraints in obtaining fine-scale canopy parameters, including low image resolution, significant weather impact, and diminished real-time capabilities, making it challenging to meet the demands for individual tree canopy segmentation and height estimation [13], [17], [18], [19]. In contrast, images captured by unmanned aerial vehicles (UAVs) offer advantages such as high resolution, low cost, and ease of data acquisition. Furthermore, high-precision canopy point clouds with texture information can be generated from overlapping UAV images through aerial triangulation and dense matching techniques. These advantages can, to some extent, compensate for the shortcomings of LiDAR and satellite imagery in estimating canopy height in complex mountainous terrains, fulfilling the requirements for large-scale and fine-grained extraction of fruit tree canopy height [20], [21], [22]. Consequently, data derived from UAV photogrammetry (UAVPD) can serve as a reliable source for the extraction of citrus tree information. This study also focuses on estimating the canopy height of fruit trees using UAVPD.

In tree canopy extraction tasks, traditional algorithms such as watershed algorithms, edge detection, and threshold segmentation are popular due to the simplicity of data processing [23]. However, detecting individual trees in complex scenes in mountainous areas is still a challenge for traditional methods due to the overlapping of tree canopies and large differences in shape and size [24]. In addition, traditional methods are unable to deeply mine the texture information of tree canopy, which limits the improvement of canopy extraction accuracy and does not solve the problem of tree height estimation. Despite UAVPD's capability to capture fine surface details of citrus tree plantations, the undulating terrain in these areas makes it difficult for traditional methods to represent such complex nonlinear variations with explicit functional relationships. This poses significant challenges for the estimation of citrus tree canopy height.

In recent years, deep learning has achieved significant success in the domain of image processing [25], and due to its superior performance in nonlinear representation, it has been extensively applied to tree canopy segmentation [26], [27], [28] and canopy height estimation in forestry [23], [29], [30], [31]. Typically, existing methods for canopy semantic segmentation utilize high-resolution true-color orthophotos, generated from UAVPD, as network inputs and employ convolutional neural network (CNN)-based architectures to extract canopy features [32], [33]. Furthermore, canopy height is estimated through the use of three-dimensional dense point clouds or canopy height models (CHMs) derived from UAVPD. The crux of these methods lies in the network's ability to extract abundant spatial information (e.g., position, shape) and semantic information (e.g., class, attributes) from true-color images, thereby enabling effective canopy extraction in forest environments with uniform tree structures, especially in flat terrains. However, deep learning methods that rely solely on true-color imagery have certain limitations: 1) they cannot capture the height features of tree canopies, making them less suitable for canopy height extraction tasks; and 2) in complex mountainous plantation areas, the undulating terrain complicates the differentiation of features with spatial characteristics similar to those of citrus tree canopies, thus constraining further improvements in the accuracy of deep learning-based canopy extraction. To address these limitations, some researchers have fused various data types, such as visible spectral imagery, multispectral imagery, digital surface models (DSMs), and CHMs, through channel stacking. They then apply deep learning methods to extract semantic features from the fused data sources [22], [34], [35]. These methods help leverage features from other data sources to enhance the distinction between tree canopies and the background, while also introducing additional feature information of tree canopies, such as CHMs, to obtain additional biomass indicators of tree canopies, such as tree height. Under typical circumstances, it is evident that the addition of multiple data sources brings forth more supplementary information, which aids in enhancing the accuracy of fruit tree extraction. For example, Hao et al. [22] extracted large-scale plantation canopy heights from UAV imagery and field-surveyed canopy height data using CNN-based methods. Xie et al. [36] achieved a correlation coefficient of 0.871 between the segmentation results and the ground truth by weighting and fusing RGB images with the CHM and inputting them into a Mask R-CNN network for canopy segmentation.

Although the above studies utilized deep learning-based multisource data combination methods to mitigate the impact of terrain undulation on canopy extraction and tree height estimation to a certain extent. However, on one hand, these methods destroy the original data structure through channel combination, making it difficult for the network to extract complementary features from different data sources and hindering the co-evolution of different data features. On the other hand, the fused multichannel data can experience severe interference from different modality features during the single-channel network feature extraction process, which inhibits the network's ability to effectively extract high-quality canopy height features from data sources such as CHM and DSM. Moreover, for the task of canopy height extraction, existing deep learning methods heavily rely on the accuracy of CHM, which is highly dependent on the accuracy of the digital terrain model (DTM). Traditional methods for obtaining DTM struggle to account for the complex nonlinear variations in vegetated areas. Therefore, it is an urgent problem to construct a method applicable to canopy segmentation and tree height estimation in complex scenes in mountainous areas by digging deeply into the canopy and tree height features directly from RGB and DSM images. In this study, we innovatively introduce the idea of "co-evolution" and explore a geometry and attributes co-evolutionary network of multisource data to achieve deep interaction and calibration of multisource data, so as to achieve the goal of high-precision citrus tree canopy segmentation and tree height estimation in complex terrain. This study fills the research gap in this field by using the co-evolutionary approach to mine the tree height information from RGB and DSM data and provides important support for the accurate extraction of citrus canopy height using DTMs and true-color images in complex mountainous terrains.

In this study, we propose a geometric and attribute coevolutionary network (GACNet) for multisource data feature interaction and aggregation, which independently extracts high-quality features from true-color imagery and other data sources, designed for the extraction of citrus tree canopy height from UAVPD in complex terrains. In detail, GACNet comprises two core modules: 1) Multisource feature interaction module (MFIM): achieves deep interaction of geometric and attribute features from different data sources by leveraging correlations in location, channel, and spatial dimensions. This allows the network's multiple data streams to focus more on each other's complementary features, and to calibrate feature responses from different data streams, thereby reducing feature discrepancies between different data sources. 2) Multisource feature aggregation module (MFAM): employs two depthwise separable convolution layers to aggregate features from multiple independent data streams. A soft attention mechanism is utilized to efficiently extract tree canopy features from the two data streams, thereby enhancing the model's output features. Additionally, in the decoding phase of the GACNet, to enhance the network's capability to discern small-scale canopies within complex terrains and to bolster its generalization across various terrain scenarios, this study has devised a multilevel feature aggregation decoder (MFA-Decoder) that is founded on cross-attention mechanisms.

The main contributions of this study are as follows:

1) We propose a GACNet, which enables effective large-scale segmentation and height extraction of citrus trees in complex mountainous terrain scenarios using only UAVPD, such as true-color orthophotos and DSMs.
2) We designed the MFIM and MFAM modules for the calibration and aggregation of multisource data features (e.g., geometric and attribute features). Furthermore, we have designed the MFA-Decoder, which significantly enhances the capability to discern small-scale canopies while markedly reducing the computational expense of the model.
3) Extensive experimental validation in four diverse terrain scenarios demonstrates that our proposed model surpasses other state-of-the-art networks in terms of accuracy and computational load. Our model facilitates precise and rapid segmentation of citrus tree canopies and estimation of tree height, thereby advancing studies in the monitoring of fruit tree areas, numbers, and heights.

II. Materials and Methods

A. Study Area

The study has selected four representative areas located in Jiangxi Province, southern China, as depicted in Fig. 1. The study area, nestled within a mountainous zone, exhibits significant topographic variation, and the surrounding vegetation is lush, primarily consisting of extensive citrus orchards. Furthermore, to compare and analyze the applicability of the GACNet under various terrain and growth conditions, experiments were conducted across four representative plots. As shown in Fig. 1, these plots exhibit varying degrees of topographic relief and tree density. In fact, for citrus trees as an economic crop, fruit-bearing trees are typically around 2.5 m tall [37]. This study primarily focuses on citrus trees that have reached an economic scale, with tree heights predominantly ranging from 1 to 4 m.

B. UAV Image Acquisition and Processing

This study employed a DJI Phantom 4 RTK UAV equipped with an RGB camera for capturing high-resolution imagery. The imagery was acquired in October 2022, with data collection occurring between 10:00 AM and 2:00 PM. Then, the raw images captured by the UAV were processed using Agisoft Photoscan Professional software, and digital orthophotos maps and DSM were generated with a ground resolution of 1.5 cm/pixel through aerial triangulation.

1) Manual Annotation of Training Samples: We utilized ArcGIS 10.8 software in conjunction with RGB orthophotos to manually delineate tree canopies, resulting in a total of 2392 tree canopy annotations. Subsequently, the heights of the 2392 trees were categorized at 0.1-m intervals. For instance, citrus trees with heights ranging from 2.00 to 2.10 m were assigned a height category of 2.05 m, which was then integrated into the attribute data for each individual tree. The distribution of tree heights for the 2392 citrus trees is illustrated in Fig. 2.

In this study, we introduced a methodology for ascertaining reference values for tree height, predicated on multiview manual optimization, hereinafter referred to as MMO. Specifically, leveraging an optimized high-precision point cloud derived from ground control points, we applied the photogrammetric collinearity equations to project the initial three-dimensional coordinates of tree vertices onto multiview imagery, as illustrated in Fig. 3. The tree vertex p' was projected onto the corresponding image points p_1, p_2, and p_3 in image 1, image 2, and image 3, respectively. Thereafter, the positions of p_1, p_2, and p_3 were manually refined to attain more precise coordinates. Ultimately, the refined coordinates for the tree vertex p' were calculated through multiview forward intersection coupled with the least squares method. In a similar vein, the coordinates for three exemplary ground points, g'_1, g'_2, and g'_3, were also refined. Ultimately, (1) was employed to compute the exact heights of the trees

$$\text{TreeHeight} = p' - \frac{1}{3}\left(g'_1 + g'_2 + g'_3\right). \tag{1}$$

In the literature [38], the UAVPD horizontal accuracy is reported as 0.035 m, and the vertical accuracy as 0.048 m. Within the study area, we selected a subset of samples and conducted a comparative analysis with field measurements. Benefiting from the higher resolution of the UAV imagery we obtained, we achieved improved surveying accuracy (namely, a horizontal accuracy of 0.015 m and a vertical accuracy of 0.02 m). Consequently, we deem the tree heights derived from the MMO method to satisfy the precision requirements for the training samples in this study and are also regarded as reference true values for comparative analysis. Subsequently, we determined 478 citrus tree heights from 2392 canopies according to the principle of random sampling to validate our method. The number of citrus trees identified across the four plots were 150, 85, 94, and 149, respectively. It is important to note that within

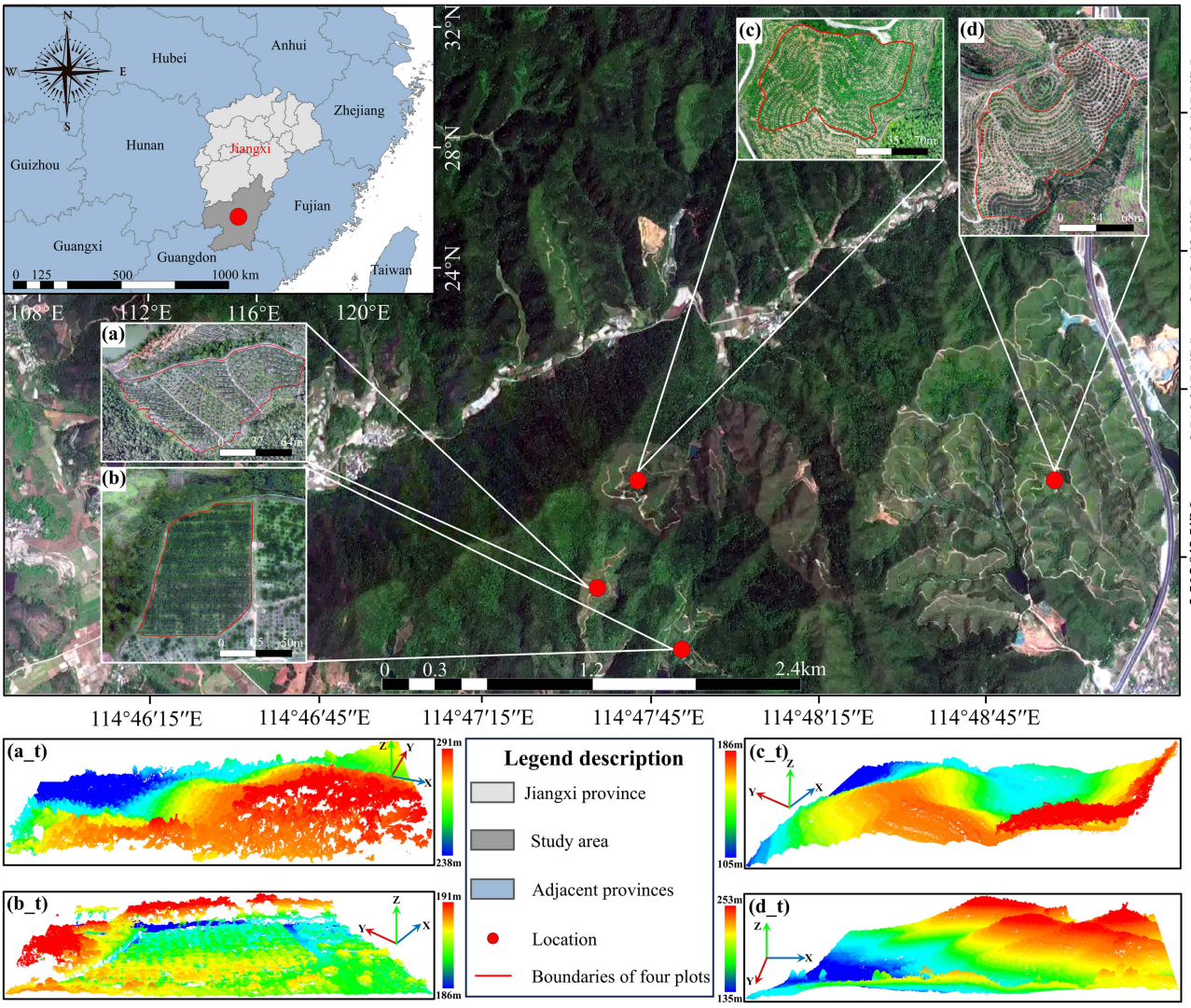


Fig. 1. Location of the study area in Xinfeng County, Ganzhou City, Jiangxi Province, China. The upper left inset in the figure represents the regional map of the study area, with the four red solid circles indicating the specific locations of the local enlarged diagrams (a), (b), (c), and (d), which correspond to the representative experimental plots. (a_t), (b_t), (c_t), and (d_t) represent the terrain variations in each of the four plots, respectively.

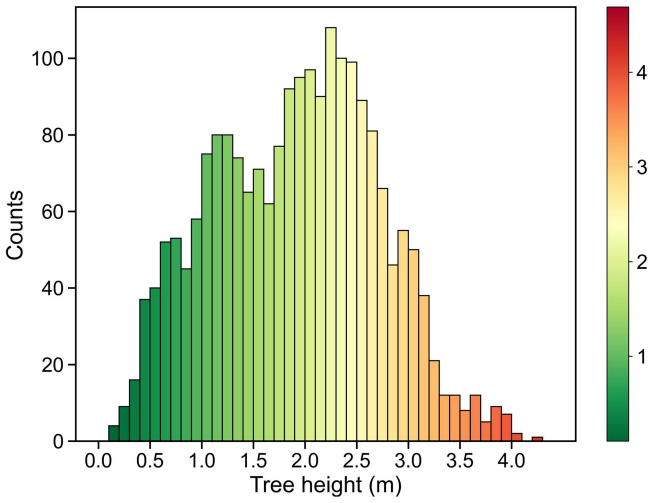


Fig. 2. Tree height distribution histogram.

the GACNet training dataset, individual trees are discriminated by the assignment of a singular height value to each tree's canopy within the annotated data; these pixel values correspond to the respective canopy heights (as illustrated in the ground truth of Fig. 4). This approach signifies that the GACNet, as proposed, has been conceptualized and designed with a dual focus on both the delineation of canopy boundaries and the extraction of canopy heights, thereby facilitating the instance segmentation of individual trees. This integrated consideration, from model architecture to sample data, is fundamental to achieving accurate instance segmentation and height extraction for solitary trees.

2) Dataset Preparation: As previously analyzed, RGB orthophotos and DSM provide foundational data for tree canopy segmentation and tree height estimation. Specifically, vegetation indices based on the visible spectrum can be used to provide initial canopy distribution. To improve the accuracy and efficiency of canopy extraction integrated with deep learning, this study utilizes vegetation indices to preselect canopy candidates.

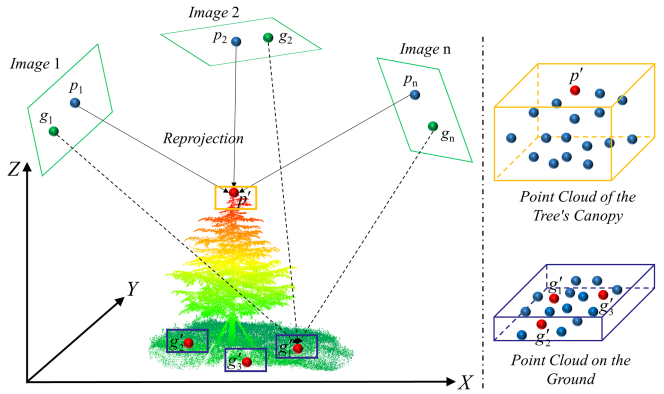


Fig. 3. Schematic representation of the calculation of tree heights based on the MMO method.

Therefore, we investigated the impact of vegetation indices such as gamma-transformed green leaf index (GGLI) [39], green leaf index (GLI) [40], and normalized green-red difference index (NGRDI) [41] on citrus canopy segmentation and height estimation under complex terrain. We computed GGLI, GLI, and NGRDI values for each plot based on RGB orthophotos using

$$\mathrm{GGLI} = 10^{\gamma}\left(\frac{2G-R-B}{2G+R+B}\right)^{\gamma} \tag{2}$$

$$\mathrm{GLI} = \frac{2G-R-B}{2G+R+B} \tag{3}$$

$$\mathrm{NGRDI} = \frac{G-R}{G+R} \tag{4}$$

where R, G, and B, respectively, represent the red, green, and blue pixel values, and γ represent the gamma value, which was set to 2.5 in this study.

This study created multimodal datasets by combining five data sources: RGB orthophotos, DSM, GGLI, GLI, and NGRDI, including combinations such as RGB-DSM, GGLI-DSM, GLI-DSM, and NGRDI-DSM, which were then used as input for GACNet. To facilitate the execution of large-scale image-based tree height extraction on a standard computer, the high-resolution experimental images were partitioned into a series of 256×256 pixel patches (i.e., each patch representing an actual area of 14.7 m^2, with an average of 5 trees per patch) for patch-wise processing. The dataset is then augmented through data enhancement techniques, including rotation, flipping, and affine transformations. Ultimately, this study procured training datasets for each data source, with each dataset comprising 12436 images. Within these images, 60% (from 1434 citrus trees) were designated for the training set, 20% (from 480 citrus trees) for the validation set, and the remaining 20% (from 478 annotated citrus trees) constituted the test set. Here, what we need to clearly state is that the validation and test sets do not participate in the model training. The RGB data, labels, and vegetation index data used in the experiments are all in PNG format, while the DSM data is stored in 32-bit TIFF format to represent tree height information.

C. Proposed Method

Our work comprises two main components: network prediction and downstream applications, as depicted in Fig. 4. Initially, we conduct canopy segmentation and tree height estimation using the proposed GACNet. The network computes the error loss by comparing the predictive outcome, denoted as Pred, with the ground truth. It then adaptively adjusts the weight parameters of the network based on this loss, yielding more precise predictive results. Subsequently, we extract the canopy from the RGB image using the acquired single-band canopy segmentation image, which also allows us to determine the number of citrus trees in the area. In addition, we derive the height of each individual citrus tree based on the canopy categories estimated by the GACNet and constructed a 3D model of the tree height of each tree.

The architecture of the GACNet is depicted in Fig. 5. The GACNet takes into account the geometric and attribute correlations between multisource data in terms of position, channel, and spatial relationships, and enhances the deep interaction and fusion of feature maps at various levels. Specifically, the GACNet employs a dual-branch parallel design to extract geometric and attribute features from multimodal data in an interactive and fused manner, with each branch aimed at capturing the unique characteristics of the corresponding input data. Given the significant differences in features between data types such as RGB and DSM, we have designed the MFIM for multisource data features between the two branches. This module facilitates cross-modality feature interaction and correction, enhancing unique characteristic extraction through a geometric and attribute co-evolution mechanism, thereby improving GACNet's capabilities in canopy segmentation and tree height estimation.

At each stage of the dual-branch architecture, the feature interaction and correction promote high-quality feature extraction in the deeper layers of the GACNet. Additionally, to extract robust features that represent citrus tree canopies and tree heights while reducing model complexity, we designed the MFAM. This module effectively integrates features from different data sources (geometric and attribute features) through a soft attention mechanism, which helps focus on key information while reducing interference from irrelevant information. Finally, we have designed the MFA-Decoder, which leverages a cross-attention mechanism to continuously interact with and aggregate geometric and attribute features from different levels, and enhance the model's perception of multiscale tree canopy. Subsequently, a lightweight multilayer perceptron (MLP) is utilized to restore the resolution of the feature maps.

As shown in Fig. 5, this architecture mainly includes three core modules: MFIM, MFAM, and MFA-Decoder.

1) Multisource Feature Interaction Module: In data derived from UAVPD, RGB orthoimages provide advantageous information for interpretation, such as the color and texture of the Earth's surface, while datasets like the DSM can offer topographic information, including surface elevation. The two-dimensional RGB imagery and the topographic height information are typically complementary; the integration of these two types of data can yield more comprehensive spatial and attribute

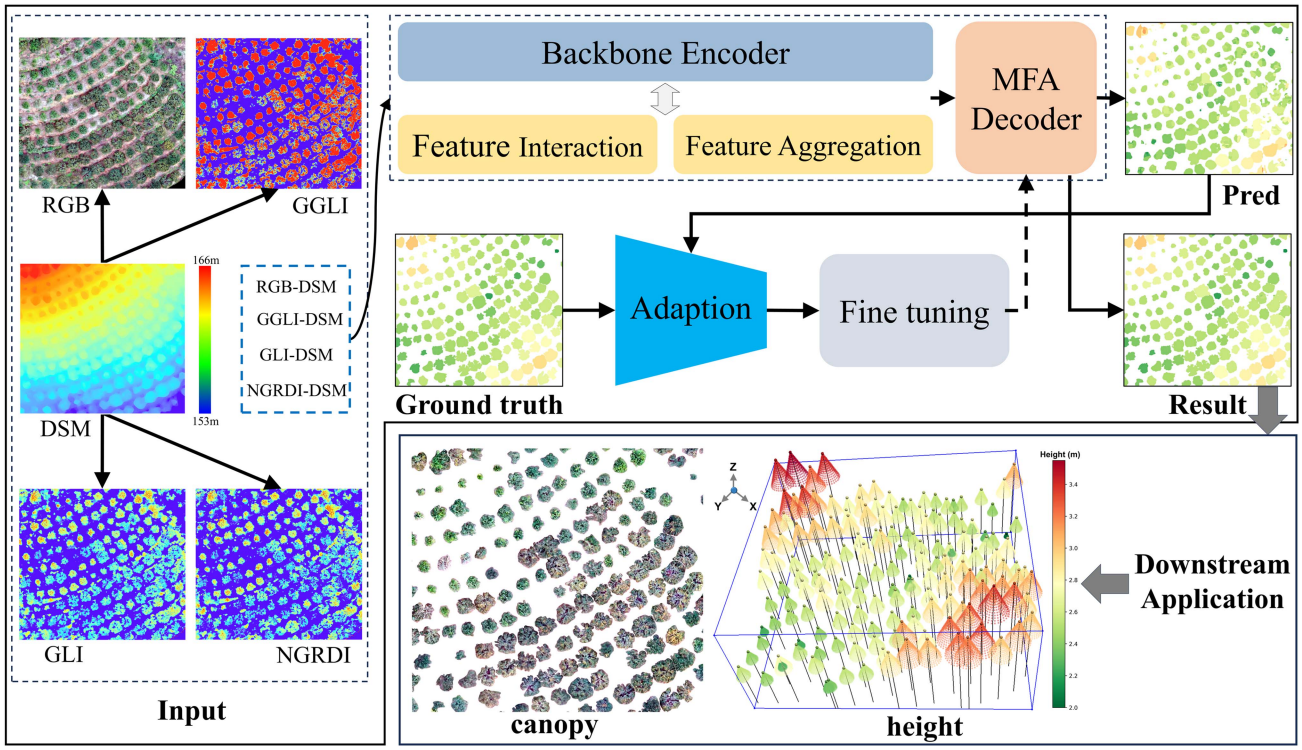


Fig. 4. Workflow of the proposed method.

information. Consequently, this study introduces the MFIM, as shown in Fig. 5(a). The MFIM provides complementary features for different data sources and achieves mutual correction of geometric and attribute features, thereby enabling the network to extract high-quality unique features from multisource data. The MFIM mainly comprises three operations: channel feature interaction (CFI), spatial feature interaction (SFI), and position feature interaction (PFI). Detailed descriptions are as follows.

a) *CFI:* To encapsulate the global information of each channel within the feature maps of the multimodal data, we initially conduct global average pooling and global max pooling operations along the channel dimension for the input multisource datasets $R_{\text{in}} \in \mathbb{R}^{H\times W\times C}$ and $X_{\text{in}} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels of the input data, respectively. Subsequently, two vectors are concatenated and an MLP is employed to facilitate deep interaction among the feature vectors of the multisource data, followed by the computation of vector weights using a sigmoid function. The resultant weights are then split into two weight vectors of equal magnitude, $W_{\text{R}}^{C} \in \mathbb{R}^{1\times 1\times C}$ and $W_{\text{X}}^{C} \in \mathbb{R}^{1\times 1\times C}$. This process can be mathematically articulated as

$$Y_{\text{GAP}}, Y_{\text{GMP}} = f_{\text{gap}}(R_{\text{in}}, X_{\text{in}}), f_{\text{gmp}}(R_{\text{in}}, X_{\text{in}}) \tag{5}$$

$$W_{\text{R}}^{\text{C}}, W_{\text{X}}^{\text{C}} = f_{\text{split}}\left(\sigma\left(f_{\text{mlp}}\left([Y_{\text{GAP}}, Y_{\text{GMP}}]\right)\right)\right) \tag{6}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, and $[,]$ denotes the concatenate operation. Through these operations, the geometric and attribute information of the two distinct modalities engages in profound interaction. Thereafter, to encourage the original feature maps to concentrate more on the tree canopy areas, we employ (7) to adjust the channel features of the original data characteristics across different modalities

$$\begin{cases} R_{\text{rec}}^{\text{C}} = W_{\text{X}}^{\text{C}} \otimes X_{\text{in}} \\ X_{\text{rec}}^{\text{C}} = W_{\text{R}}^{\text{C}} \otimes R_{\text{in}} \end{cases} \tag{7}$$

where $\otimes$ denotes channel-wise multiplication.

SFI: As the CFI is primarily utilized for the acquisition of global information from various modality data, we have further incorporated a spatial feature awareness module to facilitate the interaction and correction of local information among different modality data. Initially, the input multisource features R_{in} and X_{in} are concatenated along the channel axis, and two 1×1 convolutional layers are employed to enable the interaction of multisource features, culminating in the generation of feature map $Y_{\text{Conv}} \in \mathbb{R}^{H\times W\times 2}$. Subsequently, the weight map of the interactive features is computed using a sigmoid function and then split into two weight maps of equivalent dimensions, $W_{\text{R}}^{\text{S}} \in \mathbb{R}^{H\times W}$ and $W_{\text{X}}^{S} \in \mathbb{R}^{H\times W}$. The mathematical expression is delineated as follows:

$$Y_{\text{Conv}} = \text{Conv}\left(\text{ReLU}\left(\text{Conv}\left([R_{\text{in}}, X_{\text{in}}]\right)\right)\right) \tag{8}$$

$$W_{\text{R}}^{\text{S}}, W_{\text{X}}^{\text{S}} = f_{\text{split}}\left(\sigma\left(Y_{\text{Conv}}\right)\right). \tag{9}$$

Similar to the channel feature correction module, the spatial feature correction can be mathematically executed

$$\begin{cases} R_{\text{rec}}^{\text{S}} = W_{\text{X}}^{\text{S}} \otimes X_{\text{in}} \\ X_{\text{rec}}^{\text{S}} = W_{\text{R}}^{\text{S}} \otimes R_{\text{in}} \end{cases}. \tag{10}$$

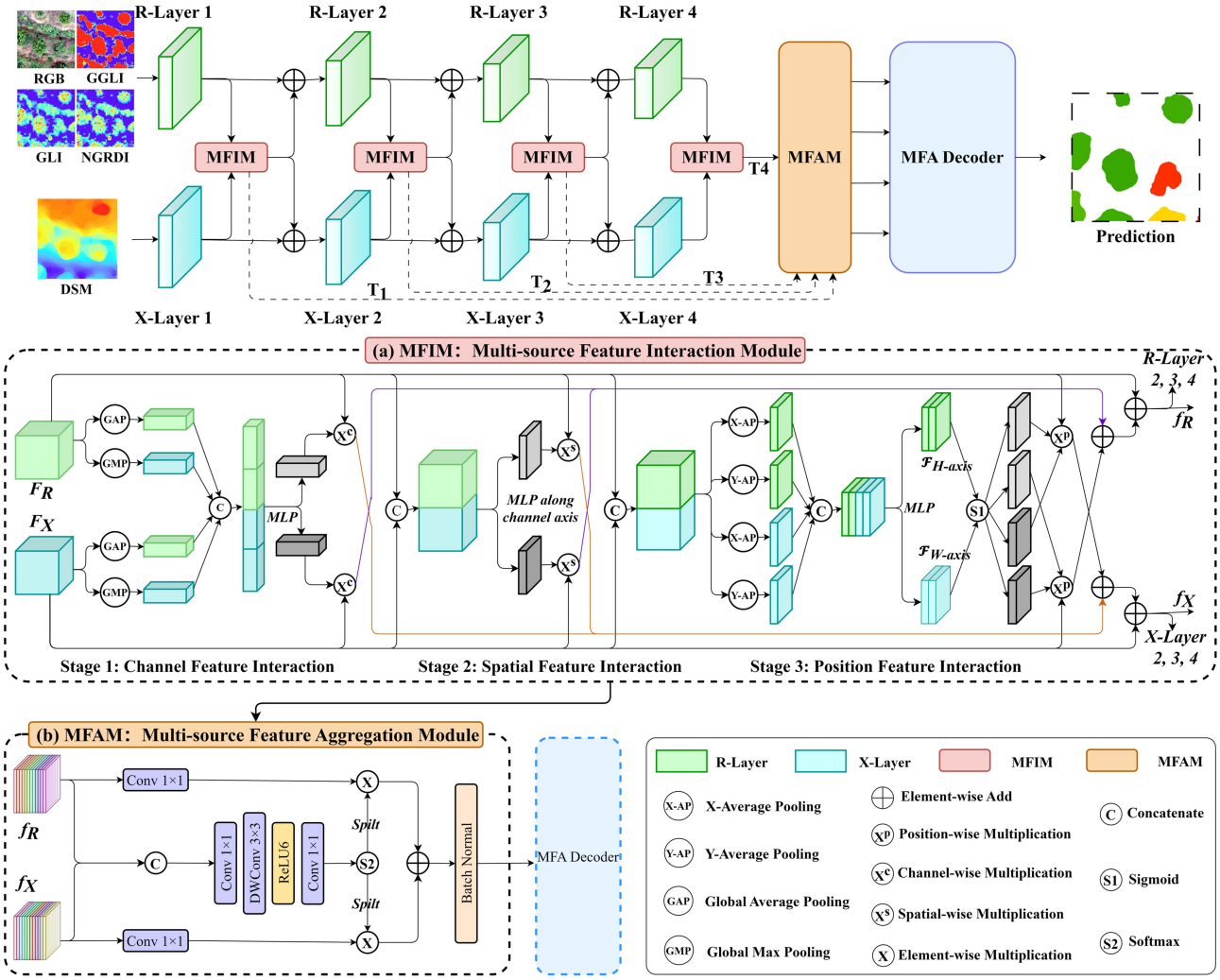


Fig. 5. Architecture of GACNet.

PFI: Due to SFI's focus on the overall spatial arrangement of feature maps rather than specific detailed locations, we have introduced a PFI mechanism to further enhance the network's ability to model long-range dependencies, thereby preserving more accurate target location information in the spatial domain. This mechanism embeds two datasets, R_{in} and X_{in}, along the horizontal (h-axis) and vertical (w-axis) directions, respectively, into four attention weight maps such as $W_{\text{R}}^{\text{h - axis}} \in \mathbb{R}^{1\times W\times C}$, $W_{\text{R}}^{\text{w - axis}} \in \mathbb{R}^{H\times 1\times C}$, $W_{\text{X}}^{\text{h - axis}} \in \mathbb{R}^{1\times W\times C}$, and $W_{\text{X}}^{\text{w - axis}} \in \mathbb{R}^{H\times 1\times C}$. Specifically, we first perform global average pooling on R_{in} and X_{in} in the horizontal and vertical directions, respectively, to retain more target location information. We then concatenate the resulting horizontal and vertical feature maps to obtain feature map $Y_{\text{cat}} \in \mathbb{R}^{H\times 1\times 4C}$. Subsequently, we employ an MLP to enable information interaction within the Y_{cat} features and use the split function to separate features in the X and Y directions, resulting in $Y_{\text{h}-axis} \in \mathbb{R}^{H\times 1\times 2C}$ and $Y_{\text{w}-\text{axis}} \in \mathbb{R}^{1\times W\times 2C}$. Finally, according to (12), we use a sigmoid function to redistribute the weights of the feature maps and then separate them into horizontal and vertical attention weights using split operation

$$Y_{\text{h}-\text{axis}}, Y_{\text{w}-\text{axis}} = f_{\text{split}}(f_{\text{mlp}}(Y_{\text{cat}})) \tag{11}$$

$$\begin{cases} W_{\text{R}}^{\text{h}-\text{axis}}, W_{\text{X}}^{\text{h}-\text{axis}} = f_{\text{split}}\left(\sigma\left(Y_{\text{h}-\text{axis}}\right)\right) \\ W_{\text{R}}^{\text{w}-\text{axis}}, W_{\text{X}}^{\text{w}-\text{axis}} = f_{\text{split}}\left(\sigma\left(Y_{\text{w}-\text{axis}}\right)\right) \end{cases}. \tag{12}$$

Next, we perform position feature correction on the original data features from different modalities using

$$\begin{cases} R_{\text{rec}}^{\text{P}} = W_{\text{X}}^{\text{h}-\text{axis}} \otimes W_{\text{X}}^{\text{w}-\text{axis}} \otimes X_{\text{in}} \\ X_{\text{rec}}^{\text{P}} = W_{\text{R}}^{\text{h}-\text{axis}} \otimes W_{\text{R}}^{\text{w}-\text{axis}} \otimes R_{\text{in}} \end{cases}. \tag{13}$$

After the operations of interaction and correction of channel, spatial, and positional information, the complete corrected features of the two modalities can be represented as

$$R_{\text{out}} = R_{\text{in}} + \lambda R_{\text{rec}}^{\text{C}} + \delta R_{\text{rec}}^{\text{S}} + \varepsilon R_{\text{rec}}^{\text{P}} \tag{14}$$

$$X_{\text{out}} = X_{\text{in}} + \lambda X_{\text{rec}}^{\text{C}} + \delta X_{\text{rec}}^{\text{S}} + \varepsilon X_{\text{rec}}^{\text{P}} \tag{15}$$

where λ, δ, and ε are hyperparameters. Experiments conducted in Section III-D were designed to ascertain the optimal values for these three parameters.

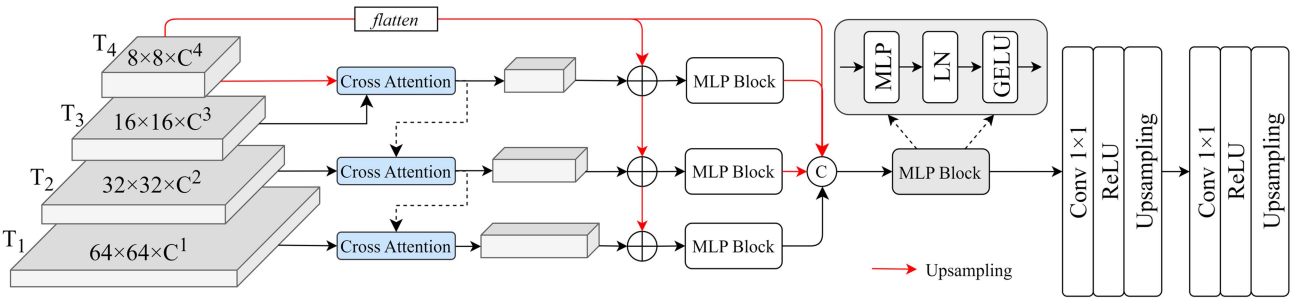


Fig. 6. Structure of MFA-Decoder.

2) Multisource Feature Aggregation Module: To enhance the aggregation of the corrected features within the MFIM, we have constructed a MFAM, as depicted in Fig. 5(b). Initially, we integrate features using the Concatenate operation, followed by the application of two depthwise separable convolutional layers to deeply integrate the characteristics of features R_{out} and X_{out}. Subsequently, the two fused features are concatenated, and the weight maps $W_{\text{R}}^{\text{FAM}} \in \mathbb{R}^{H\times W\times C}$ and $W_{\text{X}}^{\text{FAM}} \in \mathbb{R}^{H\times W\times C}$ for the two distinct features are computed using the Softmax function. Finally, features R_{out} and X_{out} are recalibrated individually using the weight maps to obtain the final fused features for feature decoding. The mathematical operation of the feature aggregation can be expressed as

$$Y_{\text{Conv}}^{\text{R}}, Y_{\text{Conv}}^{\text{X}} = \text{Conv}\left(R_{\text{out}}\right), \text{Conv}\left(X_{\text{out}}\right) \tag{16}$$

$$W_{\text{R}}^{\text{FAM}}, W_{\text{X}}^{\text{FAM}} = \text{Softmax}\left(\text{Conv}\left(\text{DWConv}\left(\text{Conv}\left(\left[Y_{\text{Conv}}^{\text{R}}, \text{X}_{\text{Conv}}^{\text{X}}\right)\right)\right)\right) \tag{17}$$

$$Y_{\text{out}} = W_{\text{R}}^{\text{FAM}} \otimes F_{\text{Conv}}^{\text{R}} + W_{\text{X}}^{\text{FAM}} \otimes F_{\text{Conv}}^{\text{X}}. \tag{18}$$

3) Multilevel Feature Aggregation Decoder: Under normal circumstances, within citrus cultivation areas, there is a presence of citrus seedlings with smaller canopy sizes and lower tree heights, which constitute a complex scenario that diminishes the recognition capability of tree canopies by deep learning networks. To address the significant scale variations in canopy shape, size, and tree height within the cultivation area, in this study, we have constructed the MFA-Decoder based on cross-attention to enhance the model's multiscale target recognition capability. The network architecture of the MFA-Decoder is depicted in Fig. 6.

Unlike conventional decoders, we first utilize a cross-attention mechanism to deeply interact and aggregate feature maps across different scales before proceeding with a layer-by-layer upsampling operation, aiming to maintain consistency in contextual information. Specifically, as shown in Fig. 7, the MFAM outputs feature maps at four scales such as $\mathbf{F}_{\text{T1}}$, $\mathbf{F}_{\text{T2}}$, $\mathbf{F}_{\text{T3}}$, and $\mathbf{F}_{\text{T4}}$. First, we upsample $\mathbf{F}_{\text{T4}}$ to the same size as $\mathbf{F}_{\text{T3}}$, then flatten the upsampled $\mathbf{F}_{\text{T4}}$ and $\mathbf{F}_{\text{T3}}$ to match the size, and apply a linear embedding layer for linear mapping to improve the model's generalization ability. Next, we use the cross-attention mechanisms to achieve feature interaction between $\mathbf{F}_{\text{T4}}$ and $\mathbf{F}_{T3}$, resulting in features $CA_{\text{T4}} \in \mathbb{R}^{256\times C^3}$ and $CA_{\text{T3}} \in \mathbb{R}^{256\times C^3}$. The cross-attention mechanism can be mathematically expressed as

$$\begin{cases} f_{\text{T4}} = K_{\text{T4}}^{\text{T}} \times V_{\text{T4}} \\ f_{\text{T3}} = K_{\text{T3}}^{\text{T}} \times V_{\text{T3}} \end{cases} \tag{19}$$

$$\begin{cases} CA_{\text{T4}} = Q_{\text{T3}} \times \text{Softmax}(f_{\text{T4}}) \\ CA_{\text{T3}} = Q_{\text{T4}} \times \text{Softmax}(f_{\text{T3}}) \end{cases}. \tag{20}$$

Subsequently, we deploy two feedforward neural network (FFN) layers and a DWConv layer to conduct nonlinear mappings on the feature vectors. It is important to highlight that we have incorporated a depthwise separable convolutional layer within the FFN, which not only augments the model's capability to articulate features across various scales but also fortifies the local interconnections within multiscale features. Thereafter, we employ a depthwise separable convolutional layer of 3×3 dimensions, coupled with two linear embedding layers, to further amalgamate the features outputted by the FFN layers while conserving a more substantial amount of spatial information, yielding a feature map $Z_{\text{out}} \in \mathbb{R}^{256\times C^3}$ that is congruent in scale with $\mathbf{F}_{\text{T3}}$. The FFN can be delineated as

$$\text{FFN} = \text{Linear}\left(\text{GELU}\left(\text{DWConv}\left(\text{Linear}\left(CA\right)\right)\right)\right). \tag{21}$$

Finally, we aggregate $\mathbf{F}_{\text{T3}}$ and $\mathbf{F}_{\text{T4}}$ to obtain a two-dimensional vector Z_{out}, which is then resized to match the size of $\mathbf{F}_{\text{T3}}$, resulting in the final aggregated feature $CA_{\text{T4}-\text{T3}}$. In addition, we continue to upsample $CA_{\text{T4}-\text{T3}}$ to match the size of $\mathbf{F}_{\text{T2}}$ and perform feature aggregation with $\mathbf{F}_{\text{T2}}$ using the aforementioned method. Subsequently, through these iterative operations, we ultimately obtain three aggregated features, i.e., $CA_{\text{T4}-\text{T3}}$, $CA_{\text{T3}-\text{T2}}$, and $CA_{\text{T2}-\text{T1}}$.

As depicted in Fig. 6, during the layer-by-layer upsampling of the aggregated features, we utilize multiple MLP blocks to upsample the aggregated features, restoring the feature maps to the dimensions equivalent to the original input data. This approach serves to diminish the number of parameters and concurrently augment computational efficiency.

D. Evaluation Metrics

We adopt three evaluation metrics, namely overall accuracy (OA), F1 score, and mean Intersection over Union (mIoU), to evaluate the performance of our methodology in estimating the height of citrus trees under complex terrain. These three

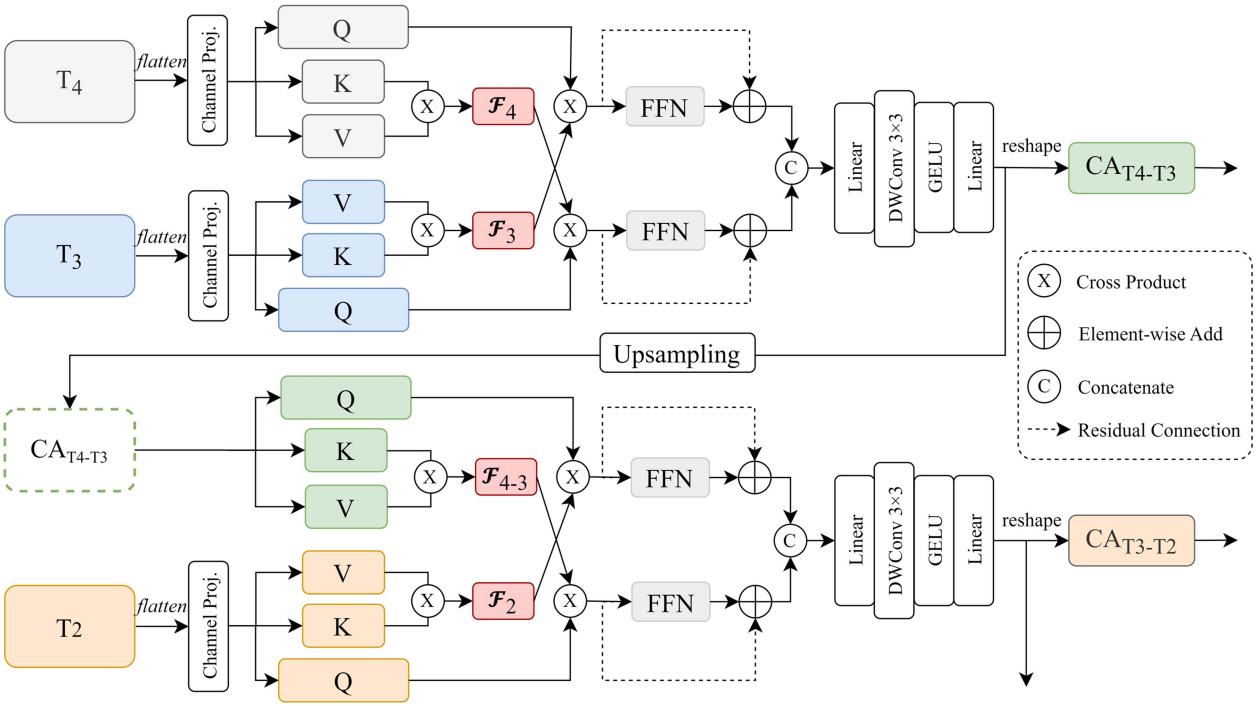


Fig. 7. Structure of multilevel feature cross-attention calculation module.

evaluation metrics can be mathematically formulated as

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{22}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{23}$$

$$\mathrm{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{24}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. k denotes the number of classification categories. Precision and recall can be calculated as

$$\mathrm{Precision\ (P)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{25}$$

$$\mathrm{Recall\ (R)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{26}$$

In addition, we also employ MAE, RMSE, and the coefficient of determination (R^2) as metrics to evaluate the performance of our method in canopy height extraction. The rationale for choosing these three metrics is as follows: MAE quantifies the average absolute error between true and predicted canopy heights in the test dataset, while RMSE indicates the standard deviation of residuals (prediction errors); lower MAE and RMSE values indicate better performance in canopy height extraction. Furthermore, R^2 assesses the goodness of fit of the model to the canopy height data; values closer to 1 indicate better fit. Therefore, these three metrics comprehensively evaluate the performance of our method. The calculation formulas for MAE, RMSE, and R^2 are provided below

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=0}^{N} |y_i - \hat{y}_i| \tag{27}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2} \tag{28}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \tag{29}$$

where N is the number of canopy height extracted by the model, y_i is the true canopy height, $\hat{y}_i$ is the predicted canopy height, and $\bar{y}_i$ is the average height from the assessed datasets.

E. Implementation Details

All experiments in this study were conducted using the PyTorch deep learning framework on a single computer equipped with an NVIDIA GeForce RTX 3060 GPU (12 GB video memory). All models were trained using the AdamW optimizer with a weight decay rate of 0.001, an initial learning rate of 1e-4, a batch size of 8, and 50 epochs. We employed the Poly learning rate decay strategy and used cross-entropy as the model's loss function. The two branch backbone networks of the proposed GACNet utilized the encoder (MiT) from SegFormer [42], which comprises four stages of Transformer modules, each containing a varying number of transformer layers. To expedite convergence on our dataset and reduce dependence on large-scale datasets, the MiT backbone network was pretrained on ImageNet.

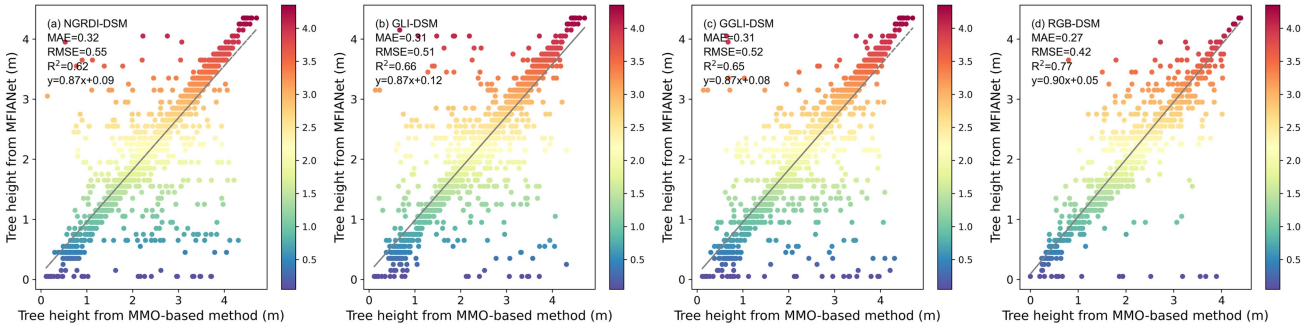


Fig. 8. Linear regressions between reference values and GACNet estimates. (a)–(d) Show the linear regressions of NGRDI-DSM, GLI-DSM, GGLI-DSM, and RGB-DSM as inputs, respectively.

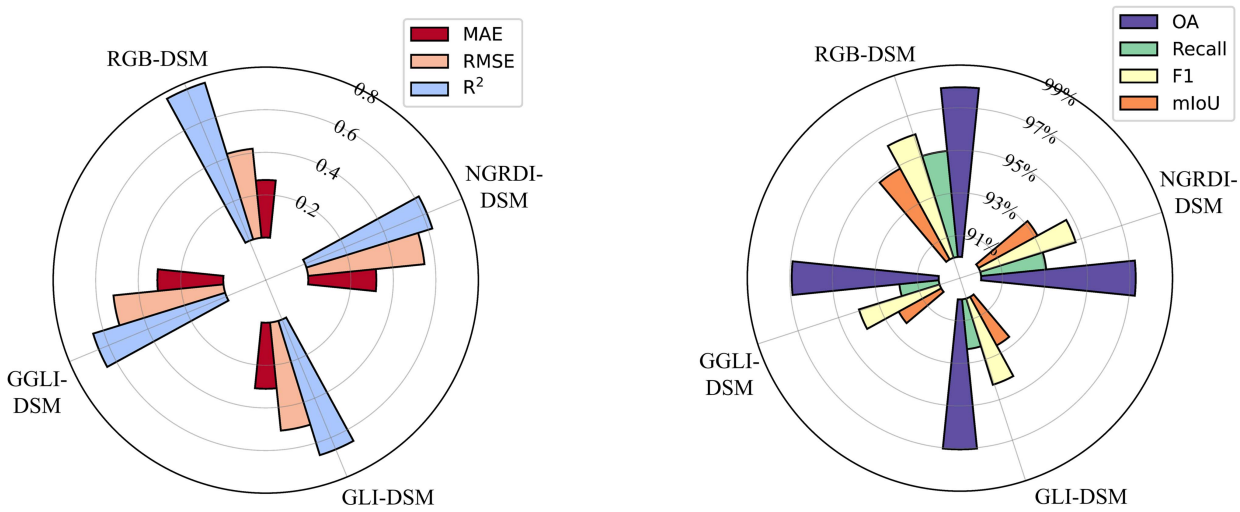


Fig. 9. Comparison of citrus canopy segmentation and tree height estimation outcomes utilizing diverse data sources.

III. Experimental Results

In this section, we first evaluated and analyzed the impact of different data sources on citrus canopy segmentation and height estimation using the GACNet. Then, we compared and analyzed the GACNet with classical and state-of-the-art networks to validate the superiority of the GACNet. Finally, we conducted ablation experiments to verify the effectiveness and rationality of the different module designs within the GACNet.

A. Citrus Tree Height Estimation

To validate the efficacy of the GACNet, we utilized a test dataset comprising 478 citrus trees to assess the performance of canopy height extraction. The tree heights estimated by the GACNet (MiT-B2) were compared and analyzed against those obtained using the MMO-based method. The accuracy of tree height estimation by the GACNet, with inputs of RGB-DSM, GGLI-DSM, GLI-DSM, and NGRDI-DSM, was evaluated.

Fig. 8 illustrates the linear regression of tree height estimation by GACNet using different modalities of input data. Referring to the citrus tree heights determined by the MMO-based method, GACNet was able to estimate the height of citrus trees in complex terrain ($R^2 > 0.60$) when using four dataset input modes, demonstrating overall consistency with the reference values. The GACNet, using RGB-DSM as input, achieved the highest accuracy in tree height estimation as depicted in Fig. 9 (MAE = 0.27, RMSE = 0.42, $R^2 = 0.77$), indicating that the GACNet can extract high-dimensional terrain and tree height features from DSM, thereby mitigating the impact of complex terrain on tree height estimation. Compared to the use of GGLI-DSM, GLI-DSM, and NGRDI-DSM, the use of RGB-DSM as input resulted in RMSE reductions of 19.2% , 17.6% , and 23.6% , respectively, and R^2 increases of 18.4% , 16.6% , and 24.1% , respectively.

As depicted in Fig. 8, it can be observed that there is a significant scattering of points in the regions below 1 m and above 4 m. This indicates that models utilizing GGLI-DSM, GLI-DSM, and NGRDI-DSM datasets encounter difficulties in detecting citrus tree canopies with heights below 1 m or above 3.5 m. Such limitations can lead to significant underestimations or overestimations of citrus tree heights and even frequent occurrences of erroneously estimating tree heights as zero. This limitation is primarily due to the scarcity of canopies below 1 m and above

TABLE I
ACCURACY OF TREE HEIGHT ESTIMATION USING THE GACNET (MIT-B2) WITH RGB-DSM AS THE INPUT ACROSS VARIOUS RANGES OF TREE HEIGHTS

Height (m)	Number of Citrus Trees	MAE (m)	RMSE (m)
1-1.5	66	0.17	0.38
1.5-2	73	0.15	0.32
2-2.5	146	**0.12**	**0.27**
2.5-3	78	0.15	0.31
3-3.5	49	0.22	0.44
3.5-4	27	0.33	0.53
>4 & <1	23 & 16	0.39	0.61

Bold numbers indicate the optimal value for this metric.

4 m, resulting in the GACNet extracting limited information regarding these canopy categories. Regarding the accuracy of tree height estimation, it is evident that the model using RGB-DSM data closely approximates the heights obtained through the MMO-based method, with a coefficient of determination (R^2) reaching 0.77. This underscores that tree heights derived from the MMO-based method serve as suitable foundational data for model training. The well-trained GACNet effectively circumvents the prerequisite for precise prior computation of the DEM and CHM to derive tree heights across large-scale areas.

To assess the impact of citrus canopy height on tree height extraction, we evaluated the accuracy of canopy height estimation within various height ranges (refer to Table I). It can be observed from Table I that the GACNet demonstrates the highest accuracy (MAE = 0.12 m, RMSE = 0.27 m) when citrus trees have heights ranging from 2 to 2.5 m. The MAE metrics for citrus trees with heights between 1 and 3 m are all below 0.2 m, indicating that the GACNet is adept at accurately obtaining tree heights for a wide spectrum of citrus trees. Conversely, the GACNet exhibits the lowest accuracy (MAE $>$ 0.35 m, RMSE $>$ 0.60 m) when canopy heights exceed 4 m or fall below 1 m. This discrepancy is mainly due to two factors: 1) Since the main subjects of this study are citrus trees that have reached an economic scale, trees below 1 m and above 4 m are relatively scarce, significantly fewer than trees in other height ranges; 2) the photogrammetry techniques used in this study are unable to penetrate the canopy, leading to potential inaccuracies in height estimation for short citrus trees due to interference from other vegetation, thereby resulting in lower accuracy for GACNet in this height range.

B. Citrus Tree Canopy Segmentation

In fact, accurate segmentation of citrus trees is a prerequisite step for estimating the heights of the individual trees, as only such precise delineation can effectively anchor the trees for height estimation. Hence, this study also validates the efficacy of the GACNet in segmenting citrus tree canopies under complex terrain. Visual assessments presented in Fig. 10 demonstrate the results of canopy segmentation and tree height estimation using RGB-DSM as the input. It is evident that the GACNet exhibits strong robustness and generalization capabilities when addressing citrus tree canopies of various sizes and shapes in complex terrains and diverse background environments.

In terms of evaluation metrics, including Recall, F1 score, and mIoU, the canopy segmentation results using the GACNet based on four data input modes are presented in Figs. 9 and 11. The outcomes obtained from the GACNet demonstrated commendable performance in canopy segmentation (Recall $>$ 91.86% , F1 score $>$ 93.97% , mIoU $>$ 92.28%), with the GACNet utilizing RGB-DSM data achieving the highest mIoU index of 94.86% . Compared to the GACNet using GGLI-DSM, GLI-DSM, and NGRDI-DSM as inputs, the mIoU for RGB-DSM data was found to be 2.58% , 2.26% , and 1.70% higher, respectively. This discrepancy primarily arises from the fact that network models employing vegetation index inputs lack the spectral and textural characteristics essential for accurate canopy segmentation, which leads to inferior performance in the recognition of citrus canopies.

Table II presents a comparative analysis of GACNet's accuracy in canopy segmentation and tree height estimation when employing backbone networks of varying scales. Two representative lightweight models were selected for comparison: EfficientNetV2 from CNNs and SegFormer (MiT) from Transformers. As demonstrated in Table II, GACNet achieved its highest canopy segmentation accuracy (mIoU) of 95.08% and 92.02% when utilizing the SegFormer-B3 (MiT-B3) and EfficientNetV2-S backbones, respectively. For the EfficientNetV2 backbone, the accuracy of GACNet's canopy segmentation declines as the network scale increases, whereas for the SegFormer backbone, accuracy initially improves but then decreases with increasing network depth. This suggests that augmenting network depth does not invariably result in enhanced extraction accuracy, likely due to the noise introduced by parameter redundancy in deeper networks, which constrains the improvement in accuracy. Notably, when EfficientNetV2 serves as the backbone, the fixed portion of the GACNet model reaches a maximum parameter count and computational complexity of 4.67M and 52.22G, respectively, leading to slower inference times. The variations in parameter count and computational complexity of the GACNet fixed portion across different backbone networks stem from differences in the number of feature map channels input to this segment. With the SegFormer backbone, the maximum parameter count and computational complexity of the fixed part of GACNet are 6.33M and 56.64G, respectively. Furthermore, the majority of GACNet's parameters are concentrated in the backbone network, with the fixed portion of the model accounting for only 14.8% of the total parameter count. This indicates that the proposed MFIM, MFIA, and MFA-Decoder modules have a relatively small parameter count, allowing GACNet to flexibly replace the backbone network to accomplish other semantic segmentation tasks without significantly increasing the computational burden.

C. Performance Comparison With Other State-of-the-Art Networks

To further validate the performance of the GACNet in the segmentation and height extraction of citrus tree canopies under

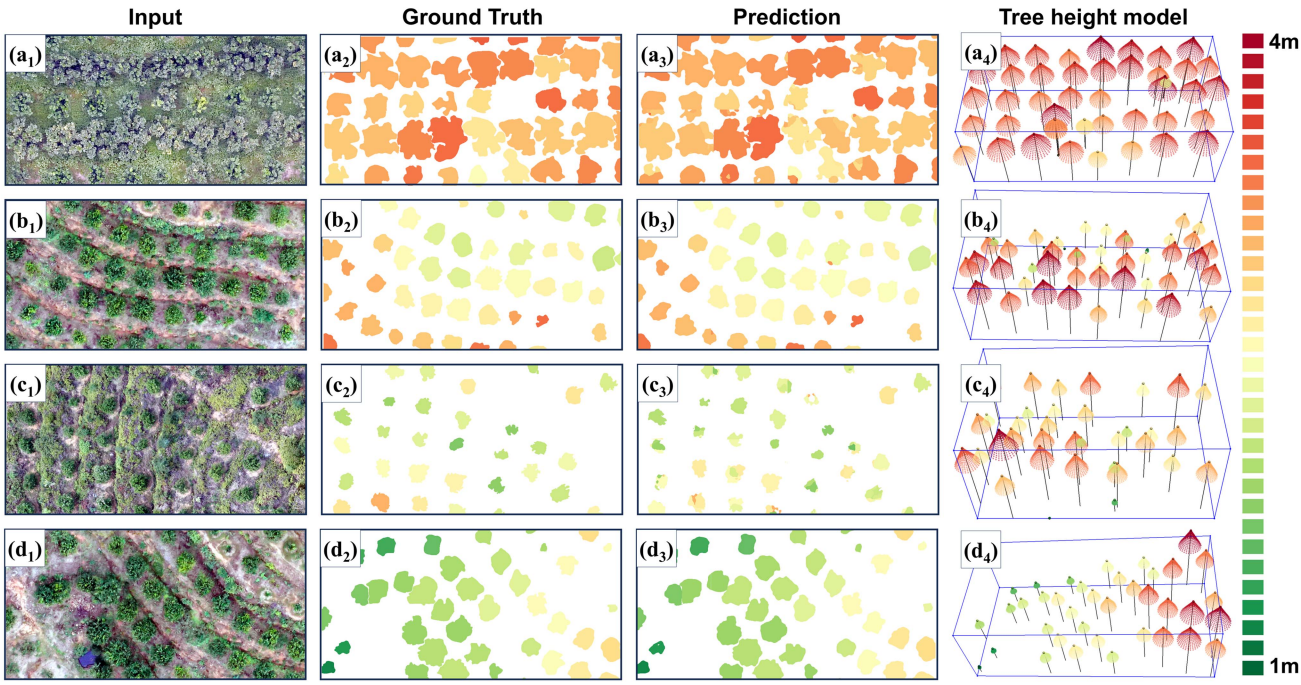


Fig. 10. Visualization outcomes of canopy segmentation and tree height estimation conducted by the GACNet (MiT-B2) with RGB-DSM as input. (a), (b), (c), and (d) correspond to four representative areas selected for comparative analysis.

TABLE II
COMPARISON OF CANOPY SEGMENTATION IN THE GACNET MODEL ACROSS DIFFERENT SCALES OF BACKBONE NETWORKS (RGB-DSM INPUT)

Backbone	F1 (%)	mIoU (%)	MAE (m)	RMSE (m)	#Params (M)	FLOPs (G)	Speed (FPS)
EfficientNetV2-S	94.05	92.02	0.34	0.59	22.75 +2.06	59.14 +46.56	13.09 -4.02
EfficientNetV2-M	93.88	91.74	0.27	0.46	112.32 +3.32	158.36 +48.35	5.26 -2.84
EfficientNetV2-L	91.23	89.76	0.33	0.61	238.72 +4.67	355.59 +52.22	4.27 -3.25
MiT-B0	92.85	91.32	0.29	0.50	**10.71 +2.33**	**11.51 +45.71**	**19.71 -8.29**
MiT-B1	95.26	94.08	0.27	0.45	42.52 +6.33	42.36 +56.64	14.30 -4.63
MiT-B2	96.03	94.86	0.27	0.42	64.61 +6.33	76.28 +56.64	9.74 -2.23
MiT-B3	**96.19**	**95.08**	**0.25**	**0.38**	104.37 +6.32	136.92 +56.64	7.10 -5.69
MiT-B4	95.91	94.69	0.26	0.42	137.91 +6.32	195.81 +56.64	4.60 -1.60
MiT-B5	95.64	93.43	0.27	0.46	179.11 +6.33	254.79 +56.56	4.10 -2.14

Bold numbers indicate the optimal value for this metric. The first half of the plus sign in the table denotes the #Params, FLOPs, and speed of the backbone network, while the second half represents the #Params, FLOPs, and speed of the fixed part of the GACNet model.

complex terrain conditions, we conducted a comparison between the GACNet and the current state-of-the-art networks, including FCN [43], BiseNetV2 [44], UNet [45], HRCNet [46], DeepLabV3+ [47], EfficientNetV2 [48], TransUNet [49], CSwin Transformer [50], SegFormer [42], ConvNeXtV2 [51], and Samba [52].

As depicted in Table III, the GACNet achieved the highest accuracy in canopy segmentation (e.g., mIoU = 95.08%), with a maximum improvement of 7.32% in mIoU compared to other representative networks. Compared to networks such as DeepLabV3+, CSwin-Base, and SegFormer-B4, which use channel-stacked data as input, the GACNet, which employs multisource data interaction, demonstrated higher accuracy in both canopy segmentation and tree height estimation, with mIoU improvements of 3.62% , 3.06% , and 2.79% , respectively, and RMSE errors reduced by 25.4% , 22.4% , and 20.8% , respectively. Furthermore, compared to the recently proposed ConvNeXtV2-Base and Samba networks, the canopy segmentation accuracy of the GACNet model was improved by 3.24% and 1.81% , and the tree height estimation error (RMSE) was reduced by 26% and 17.3% , respectively. This indicates that for these state-of-the-art networks, effectively complementing multisource features when using channel-stacked data is challenging, thus limiting the improvement in accuracy for canopy segmentation and tree height estimation. In contrast, the GACNet, by utilizing multisource data interaction, can effectively complement multisource features of geometry and attributes, thereby exhibiting superior performance.

TABLE III
COMPARISON OF TREE CANOPY SEGMENTATION AND TREE HEIGHT ESTIMATION BETWEEN THE GACNET MODEL AND STATE-OF-THE-ART NETWORKS (USING RGB-DSM DATA AS INPUT)

Method	Backbone	Modal	OA (%)	F1 (%)	mIoU (%)	MAE (m)	RMSE (m)
FCN	VGG16	Channel Stack	94.95	90.02	87.76	0.44	0.70
BiseNetV2	/	Channel Stack	95.22	90.46	88.15	0.45	0.73
UNet	/	Channel Stack	95.52	92.05	90.56	0.36	0.62
HRCNet_W48	/	Channel Stack	95.85	92.71	91.11	0.34	0.56
DeepLabV3+	ResNet-50	Channel Stack	95.96	93.05	91.46	0.30	0.51
EfficientNetV2	/	Channel Stack	95.67	92.14	90.50	0.29	0.48
TransUNet	ResNet-50	Channel Stack	96.62	93.64	91.91	0.32	0.50
Samba	/	Channel Stack	97.71	94.87	93.27	0.31	0.46
CSwin-Tiny	/	Channel Stack	96.71	93.31	91.60	0.33	0.54
CSwin-Small	/	Channel Stack	96.92	93.81	91.80	0.32	0.51
CSwin-Base	/	Channel Stack	97.01	93.99	92.02	0.30	0.49
CSwin-Large	/	Channel Stack	96.86	93.80	91.76	0.31	0.51
SegFormer_B2	MiT-B2	Channel Stack	96.83	93.01	91.61	0.33	0.52
SegFormer_B3	MiT-B3	Channel Stack	96.90	93.33	91.90	0.31	0.50
SegFormer_B4	MiT-B4	Channel Stack	96.95	93.81	92.29	0.29	0.47
ConvNeXtV2-B	/	Channel Stack	96.94	93.37	91.84	0.34	0.52
ConvNeXtV2-L	/	Channel Stack	96.77	93.24	91.73	0.34	0.53
GACNet	EfficientNetV2-S	RGB-DSM	97.09	94.05	92.02	0.34	0.59
GACNet	MiT-B3	RGB-DSM	**98.02**	**96.19**	**95.08**	**0.25**	**0.38**

Bold numbers indicate the optimal value for this metric.

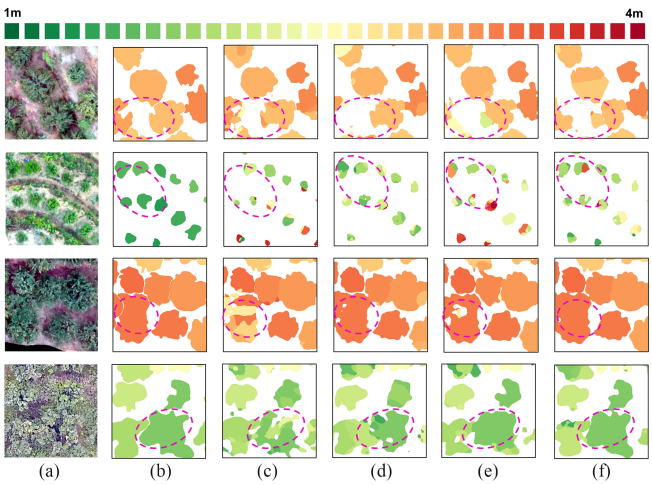


Fig. 11. Visualization results of canopy segmentation and tree height estimation for different data (all models use MiT-B2 as the backbone network). (a)–(f) Tree canopy segmentation and tree height estimation results of NGRDI-DSM, GLI-DSM, GGLI-DSM, and RGB-DSM as inputs, respectively.

The visualization results of canopy segmentation and tree height estimation for GACNet and state-of-the-art networks are shown in Fig. 12. The complex terrain and canopy backgrounds in mountainous environments impede the state-of-the-art networks from capturing the height characteristics of canopies, making it challenging to differentiate vegetation with spatial characteristics akin to those of citrus canopies. In contrast, the proposed GACNet can effectively identify the intricate edges of citrus trees and accurately estimate tree height, resulting in a complete and continuous citrus canopy with significantly reduced fragmented patches. This is primarily attributed to the MFIM and MFAM modules within the GACNet, which efficiently extract a variety of high-dimensional geometric and attribute features from RGB and DSM images, enabling deep interaction, correction, aggregation, and co-evolution of these features. This process significantly mitigates the impact of complex terrain, yielding accurate edges and tree heights for citrus canopies. Furthermore, the MFA-Decoder designed in this study effectively integrates geometric and attribute features across different scales, accurately capturing complex long-range relationships. This capability aids the GACNet in identifying small canopies. As shown in the boxes on the left side of Fig. 12(a)–(o), the GACNet accurately segments the boundaries of the canopy, including small canopies. In the boxes on the right side, the GACNet accurately estimates the height of the entire canopy, avoiding the ambiguity of multiple height values for a single tree, which other networks struggle with.

D. Ablation Study

To evaluate the efficacy of the GACNet architecture, we conducted ablation studies on the GACNet using the RGB-DSM dataset with MiT-B2 as the backbone network (refer to Table IV). The results indicate that GACNet achieved the highest accuracy in tree canopy segmentation and tree height estimation when all three modules were utilized concurrently (mIoU = 94.86% , MAE = 0.27 m, RMSE = 0.42 m). Compared to the GACNet without these modules, the mIoU increased by 3.17% , with the MAE and RMSE being reduced by 0.09 and 0.18 m, respectively. Examination of Table IV reveals that the MFIM contributes

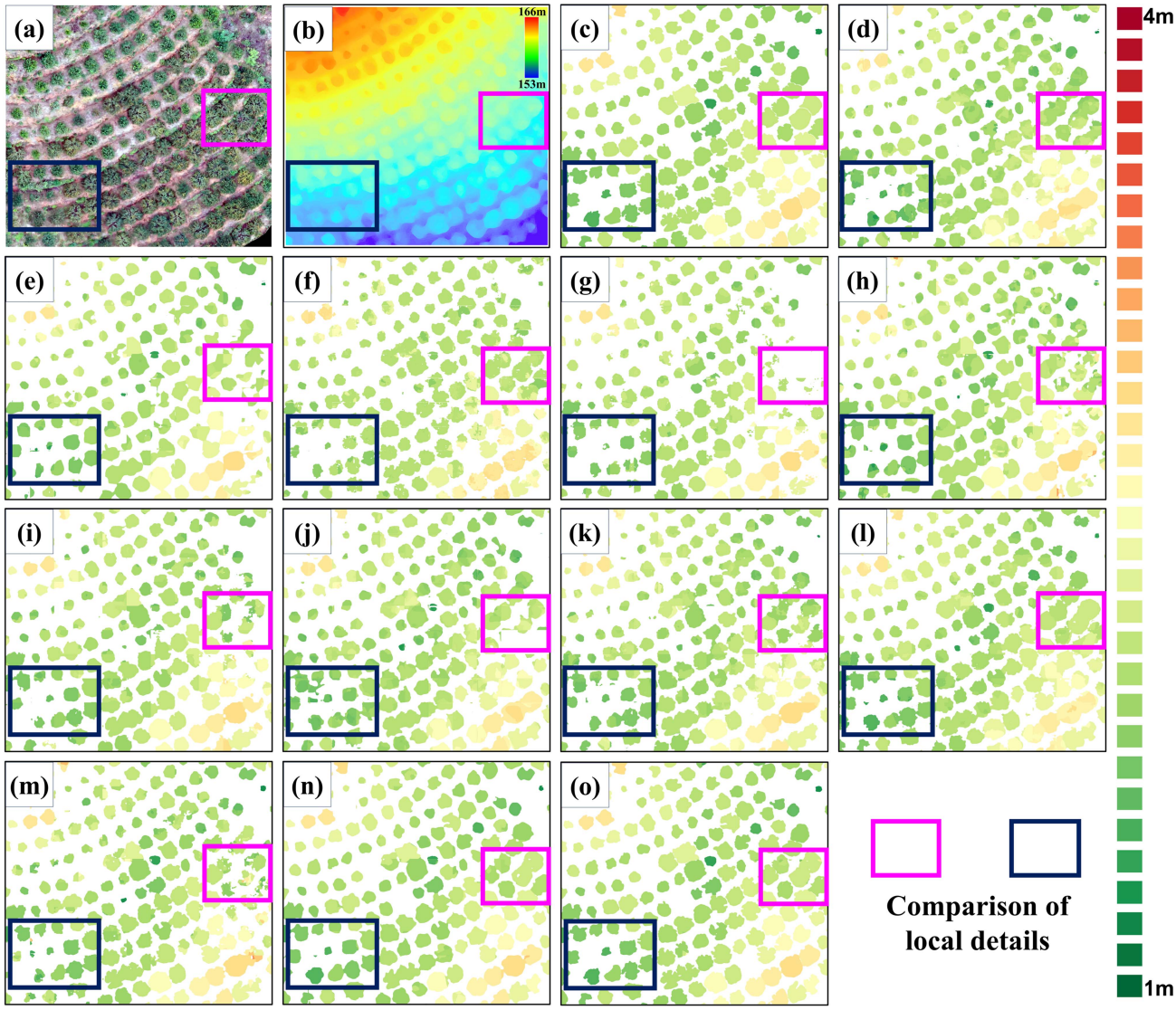


Fig. 12. Comparison of citrus canopy segmentation and tree height estimation results between GACNet and state-of-the-art networks. (a)–(o) UAV orthophotograph, DSM, ground truth, and the outcomes obtained from FCN, BiseNetV2, UNet, HRCNet_W48, DeepLabV3+, EfficientNetV2, TransUNet, CSwin-Base, SegFormer-B4, ConvNeXtV2-Base, Samba, and GACNet (MiT-B2), respectively. The rectangles are used to emphasize detail changes.

TABLE IV
ABLATION STUDY ON THE GACNET (MIT-B2) MODEL STRUCTURE (USING RGB-DSM AS INPUT)

Method	MFIM	MFAM	MFA-Decoder	mIoU (%)	MAE (m)	RMSE (m)
GACNet				91.69	0.36	0.60
	✓			93.16	0.29	0.48
		✓		92.45	0.34	0.56
			✓	92.92	0.32	0.47
	✓	✓		93.71	0.29	0.48
	✓		✓	94.24	0.29	0.47
		✓	✓	93.32	0.33	0.50
	✓	✓	✓	**94.86**	**0.27**	**0.42**

Bold numbers indicate the optimal value for this metric.

TABLE V
ABLATION STUDY OF THE CHANNEL, SPATIAL, AND POSITION INTERACTION MODULES IN THE MFIM

MFIM			mIoU (%)	MAE (m)	RMSE (m)
CFI (λ)	SFI (δ)	PFI (ε)			
0.5	0.5	0.0	94.15	0.27	0.45
0.5	0.0	0.5	94.52	0.27	0.44
0.0	0.5	0.5	94.42	0.29	0.46
0.4	0.3	0.3	94.55	0.29	0.48
0.3	0.3	0.4	94.62	0.27	0.44
0.2	0.3	0.5	**94.86**	**0.27**	**0.42**
0.2	0.2	0.6	94.75	0.29	0.47

The bold values indicate the optimal value for this metric.

the most to the model's performance, enhancing the mIoU by 1.47% and decreasing the MAE and RMSE by 0.07 and 0.14 m, respectively, when compared to the baseline model. The MFA-Decoder is the next most impactful contributor, improving the mIoU by 1.23% and reducing the MAE and RMSE by 0.04 and 0.15 m, respectively. The MFAM module's contribution is relatively modest. These improvements collectively demonstrate that the three modules developed in this study enhance the accuracy of canopy segmentation and tree height estimation in the GACNet.

We further conducted ablation experiments on the channel, spatial, and positional interaction modules within the MFIM module (refer to Table V). λ, δ, and ε represent the weight parameters for the channel, spatial, and positional features, respectively. As indicated in Table V, when these three information interaction modules are utilized concurrently, GACNet achieves the highest accuracy, thereby further validating the efficacy of the geometric and attribute multisource data feature interaction modules. When λ, δ, and ε were set to 0.2, 0.3, and 0.5, respectively, GACNet attained the optimal performance metrics (mIoU = 94.86% , MAE = 0.27 m, RMSE = 0.42 m). This outcome underscores the critical role of the interaction and correction of positional features from multisource data in enhancing GACNet's performance in tree canopy segmentation and tree height estimation.

IV. DISCUSSION

A. Large-Scale Citrus Trees Segmentation and Height Estimation

To estimate the height of the large-scale citrus trees depicted in Fig. 13(a), we segmented the image with a quarter-pixel overlap, ensuring that each tree canopy was fully captured within a single image and enhancing the completeness of the canopy segmentation results. Subsequently, the GACNet was utilized for citrus tree segmentation and height estimation within each patch. During the result merging process, we use opening operations to erode and dilate the overlapping regions to overcome the image block artifacts in the region of interest. As shown in Fig. 13(b), the GACNet is capable of accurately segmenting large areas of citrus tree canopies in complex terrain and precisely estimating the height of each individual citrus tree. The height of the citrus trees in this area is generally around 2.5 m, with significant variation in canopy size, primarily due to the presence of citrus trees at various stages of growth. Fig. 13(c)–(g) illustrates the second part of our work, which pertains to the downstream applications of the GACNet. Utilizing the outcomes of the GACNet, we can segment large-scale citrus canopies, and based on the estimated tree heights, construct three-dimensional height models of citrus trees for further analysis of extensive citrus groves, including canopy area, tree height, quantity, and density, facilitating orchard monitoring, management, and yield assessment, which holds significant practical value in precision orchard management.

Some studies [53], [54] have shown that the spatial resolution of the image directly affects the canopy detection accuracy of the model, and higher spatial resolution of the image can improve the tree detection accuracy of the deep learning model, but the resolution over a certain level does not significantly improve the accuracy of the model. Additionally, the large variation in tree canopy size in the study area is also one of the factors affecting the accuracy of the model. Based on these study results, GACNet designs a multilevel feature interaction and aggregation method for some tiny tree canopies to reduce the error caused by large differences in the size of tree canopies. Therefore, the error of GACNet model mainly comes from the spatial resolution of DSM and RGB images.

B. Model Complexity Analysis

We evaluated the spatial and temporal efficiency of the proposed GACNet model using the number of model parameters, floating point operations (FLOPs), frames per second (FPS), training time, and inference time. Table VI presents the spatial and temporal efficiency of the GACNet model compared to various mainstream networks. Despite having a higher number of parameters, the GACNet (MiT-B2) achieves lower computational complexity due to parameter sharing, and it outperforms FCN, UNet, HRCNet_W48, DeepLabV3+, EfficientNetV2, and CSwin-Base in terms of canopy segmentation accuracy.

We conducted training on an ensemble of distinct models utilizing a dataset comprising 8026 images, meticulously recording the duration of each training session. As evidenced in Table VI, the GACNet, which employs a dual-branch architecture for feature extraction, exhibits a negligible increase in training duration

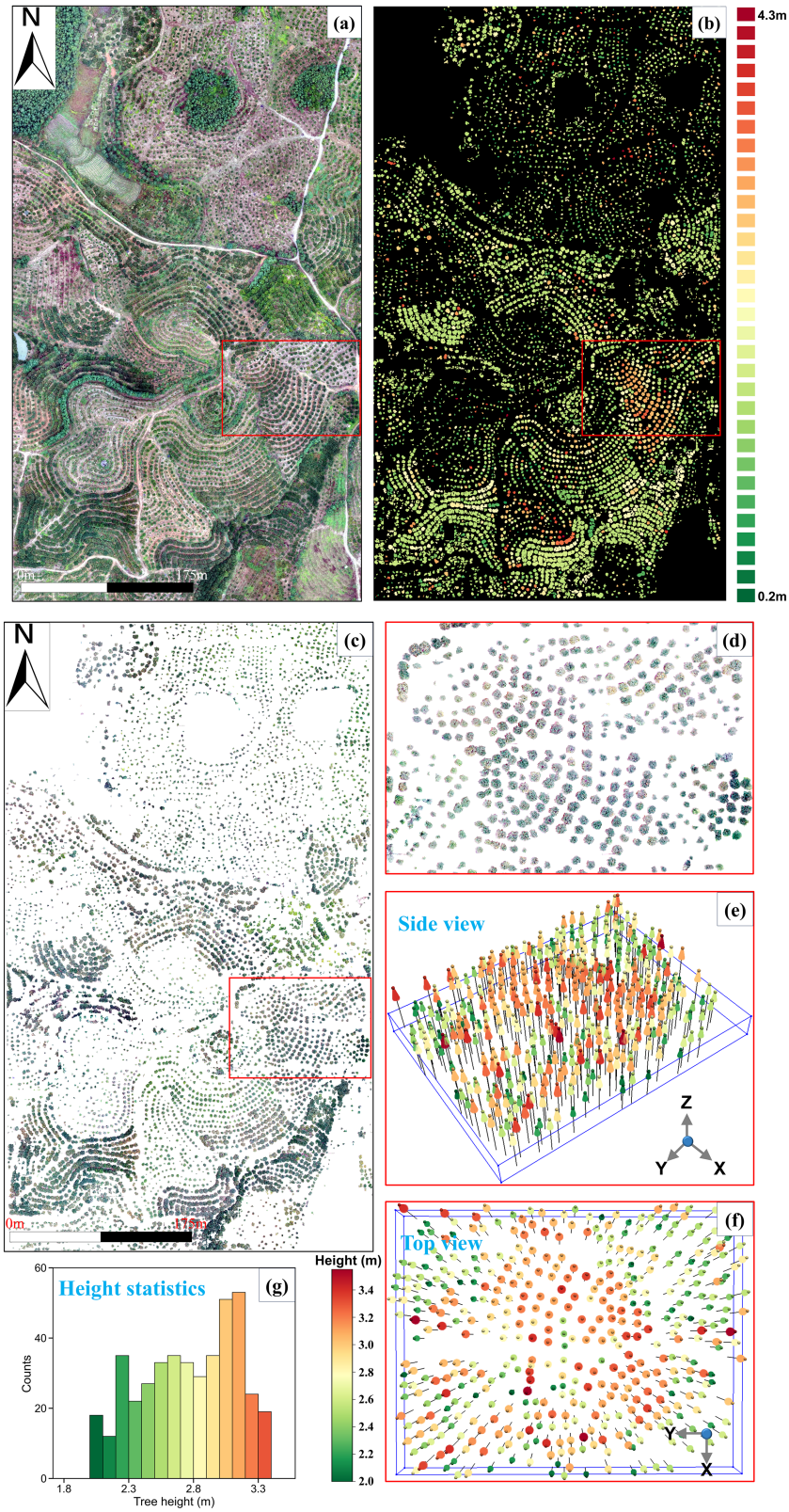


Fig. 13. Results of canopy segmentation and tree height estimation for large-scale citrus trees. (a) Large-scale citrus tree cultivation area. (b) Canopy segmentation and tree height estimation results obtained using the GACNet. (c) Ground truth for canopy segmentation. (d) A localized magnified view of the ground truth canopy segmentation. (e) and (f) Localized three-dimensional models of tree heights, respectively, constructed based on the estimated tree heights. (g) Localized tree height statistics results.

TABLE VI
COMPARISON OF THE SPACE AND TIME EFFICIENCY OF THE GACNET MODEL WITH STATE-OF-THE-ART NETWORKS (USING RGB-DSM DATA AS INPUT)

Method	Backbone	#Params (M)	FLOPs (G)	Speed (FPS)	Training time (min)	Inference time (min)
FCN	VGG16	15.90	160.97	12.93	136.16	2.66
BiseNetV2	/	**3.69**	**26.89**	9.02	**102.57**	3.81
UNet	/	13.40	248.53	8.57	196.23	4.01
HRCNet_W48	/	62.71	187.38	7.2	438.85	4.78
DeepLabV3+	ResNet-50	65.50	347.04	9.37	157.07	3.67
EfficientNetV2	/	26.76	222.62	11.89	224.93	2.89
TransUNet	ResNet-50	37.39	112.22	4.92	325.57	6.99
Samba	/	23.27	63.59	1.57	458.24	21.92
CSwin-Tiny	/	22.43	50.69	16.07	230.77	2.14
CSwin-Small	/	34.75	77.09	11.20	323.02	3.07
CSwin-Base	/	30.31	240.75	8.30	391.74	4.14
CSwin-Large	/	175.25	381.81	3.04	482.92	11.32
SegFormer_B2	MiT-B2	24.86	39.49	**16.90**	204.81	**2.03**
SegFormer_B3	MiT-B3	44.73	69.81	13.79	290.10	2.49
SegFormer_B4	MiT-B4	62.23	112.11	7.56	387.54	4.55
ConvNeXtV2-B	/	90.17	178.52	7.16	356.61	4.80
ConvNeXtV2-L	/	202.00	399.40	3.71	827.69	9.27
GACNet	EfficientNetV2-S	24.81	105.70	9.07	239.05	3.79
GACNet	EfficientNetV2-M	115.64	206.71	2.42	469.47	14.22
GACNet	EfficientNetV2-L	243.39	407.81	1.02	934.54	33.74
GACNet	MiT-B0	13.04	57.22	11.42	153.17	3.01
GACNet	MiT-B1	48.85	99.00	9.67	222.68	3.55
GACNet	MiT-B2	70.94	132.92	7.51	317.25	4.58
GACNet	MiT-B3	110.69	193.56	4.41	383.00	7.80
GACNet	MiT-B4	114.23	252.45	3.00	510.47	11.47
GACNet	MiT-B5	139.45	284.37	1.96	653.35	17.55

Bold numbers indicate the optimal value for this metric.

when juxtaposed with single-branch counterparts. This minimal escalation can be attributed to the judicious parameter sharing within GACNet's dual-branch construct, thereby curtailing the computational intricacy of the model. In comparison with the SegFormer-B3 model, which leverages MiT-B3 as its backbone, the training period for GACNet, also backed by MiT-B3, extends by an additional 92.9 min. This marginal prolongation is deemed permissible, considering the constraints imposed by the computational infrastructure employed in this study. Upon evaluating the models across a validation set of 2065 images, it was discerned that the SegFormer-B2 model outperforms in terms of inference velocity. Conversely, the GACNet (MiT-B3), although incurring an additional 5.77 min, demonstrates a marked enhancement in the precision of canopy segmentation and tree height estimation. Balancing both spatial and temporal metrics, the GACNet (MiT-B3) emerges as a model that attains superior canopy delineation and tree height estimation accuracy without substantially augmenting the computational expenditure, thus presenting a commendable attribute in the realm of practical deployment.

V. CONCLUSION

This study proposes a geometric and attribute co-evolutionary network, named GACNet, designed for the extraction of citrus tree canopies and the estimation of tree height. The impact of four data combination methods on the accuracy of canopy segmentation and tree height estimation was evaluated. The designed MFIM and MFAM modules facilitate deep interaction and mutual correction of diverse data features, effectively guiding the fusion of multimodal geometric and attribute deep features. Furthermore, an MFA-Decoder was developed, which effectively optimizes semantic information across multiple feature levels, enhancing GACNet's ability to perceive multiscale targets. Experiments conducted in four representative citrus orchards demonstrated that the proposed GACNet achieves optimal performance in citrus tree canopy segmentation and height estimation under complex terrain conditions. Compared to state-of-the-art networks, the GACNet achieved an improvement in mIoU ranging from 2% to 7% , a 13.7% reduction in MAE, and a 17.3% reduction in RMSE, with a correlation coefficient of 0.77 between tree height estimates and reference values.

Our research is capable of improving the precision and efficiency of citrus tree canopy segmentation and tree height estimation. The derived canopy and height data can be utilized to ascertain the coverage of orchards and to evaluate the growth conditions of the trees, thereby supporting the sophisticated management, conservation, and utilization of forest resources in plantations.

REFERENCES

[1] I. García-Tejero, R. Romero-Vicente, J. Jiménez-Bocanegra, G. Martínez-García, V. Durán-Zuazo, and J. Muriel-Fernández, "Response of citrus trees to deficit irrigation during different phenological periods in relation

to yield, fruit quality, and water productivity," *Agricultural Water Manage.*, vol. 97, no. 5, pp. 689–699, 2010.
[2] Y. Tang, C. Hou, S. Luo, J. Lin, Z. Yang, and W. Huang, "Effects of operation height and tree shape on droplet deposition in citrus trees using an unmanned aerial vehicle," *Comput. Electron. Agriculture*, vol. 148, pp. 1–7, 2018.
[3] H. Xu, S. Qi, X. Li, C. Gao, Y. Wei, and C. Liu, "Monitoring three-decade dynamics of citrus planting in Southeastern China using dense Landsat records," *Int. J. Appl. Earth Observ.*, vol. 103, 2021, Art. no. 102518.
[4] F. Zhang, X. Jin, J. Jiang, S. An, and Q. Lyu, "WCANet: Wavelet channel attention network for citrus variety identification," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 2845–2849.
[5] L. P. Osco et al., "Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery," *Precis. Agriculture*, vol. 22, no. 4, pp. 1171–1188, 2021.
[6] A. C. Birdal, U. Avdan, and T. Türk, "Estimating tree heights with images from an unmanned aerial vehicle," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 1144–1156, 2017.
[7] Y. Wang et al., "A novel method based on kernel density for estimating crown base height using UAV-Borne LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7004105.
[8] I. Fayad et al., "Assessment of GEDI's LiDAR data for the estimation of canopy heights and wood volume of eucalyptus plantations in Brazil," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7095–7110, 2021.
[9] Q. Liu et al., "Improving estimation of forest canopy cover by introducing loss ratio of laser pulses using airborne LiDAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 567–585, Jan. 2020.
[10] X. Huang, F. Cheng, J. Wang, P. Duan, and J. Wang, "Forest canopy height extraction method based on ICESat-2/ATLAS data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5700814.
[11] H. Zhou, J. Zhang, L. Ge, X. Yu, Y. Wang, and C. Zhang, "Research on volume prediction of single tree canopy based on three-dimensional (3D) LiDAR and clustering segmentation," *Int. J. Remote Sens.*, vol. 42, no. 2, pp. 738–755, 2021.
[12] H. Tang, H. Huang, Y. Zheng, P. Qin, Y. Xu, and S. Ding, "Improved GEDI canopy height extraction based on a simulated ground echo in topographically undulating areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5705915.
[13] Y. Liu, W. Gong, Y. Xing, X. Hu, and J. Gong, "Estimation of the forest stand mean height and aboveground biomass in Northeast China using SAR Sentinel-1B, multispectral Sentinel-2A, and DEM imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 277–289, 2019.
[14] Y. Su et al., "Spatial distribution of forest aboveground biomass in China: Estimation through combination of spaceborne lidar, optical imagery, and forest inventory data," *Remote Sens. Environ.*, vol. 173, pp. 187–199, 2016.
[15] A. Khosravipour, A. K. Skidmore, T. Wang, M. Isenburg, and K. Khoshelham, "Effect of slope on treetop detection using a LiDAR canopy height model," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 44–52, 2015.
[16] Y. Hao, F. R. A. Widagdo, X. Liu, Y. Liu, L. Dong, and F. Li, "A hierarchical region-merging algorithm for 3-D segmentation of individual trees using UAV-LiDAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5701416.
[17] H. Zhou, H. Wang, H. Song, Q. Zhang, Y. Ma, and S. Li, "Canopy height extraction over mountainous areas from GEDI lidar deconvoluted waveforms," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2502405.
[18] Q. Zhang, L. Ge, S. Hensley, G. I. Metternicht, C. Liu, and R. Zhang, "PolGAN: A deep-learning-based unsupervised forest height estimation based on the synergy of PolInSAR and LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 186, pp. 123–139, 2022.
[19] H. He, F. Zhou, Y. Xia, M. Chen, and T. Chen, "Parallel fusion neural network considering local and global semantic information for citrus tree canopy segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1535–1549, 2024.
[20] A. Matese, S. F. Di Gennaro, and A. Berton, "Assessment of a canopy height model (CHM) in a vineyard using UAV-based multispectral imaging," *Int. J. Remote Sens.*, vol. 38, no. 8–10, pp. 2150–2160, 2017.
[21] W. Li, Z. Niu, H. Chen, D. Li, M. Wu, and W. Zhao, "Remote estimation of canopy height and aboveground biomass of maize using high-resolution stereo images from a low-cost unmanned aerial vehicle system," *Ecological Indicators*, vol. 67, pp. 637–648, 2016.
[22] Z. Hao et al., "Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (mask R-CNN)," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 112–123, 2021.
[23] S. Zhang et al., "A mapping approach for eucalyptus plantations canopy and single-tree using high-resolution satellite images in Liuzhou, China," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4410413.
[24] F. H. Wagner et al., "Individual tree crown delineation in a highly diverse tropical forest using very high resolution satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 362–377, 2018.
[25] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, no. 5, Sep. 2024, Art. no. 100691.
[26] M. Aubry-Kientz et al., "Multisensor data fusion for improved segmentation of individual tree crowns in dense tropical forests," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3927–3936, 2021.
[27] J. Yang, R. Gan, B. Luo, A. Wang, S. Shi, and L. Du, "An improved method for individual tree segmentation in complex urban scene based on using multispectral LiDAR by deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6561–6576, 2024.
[28] F. Zhu, Z. Chen, H. Li, Q. Shi, and X. Liu, "CEDAnet: Individual tree segmentation in dense orchard via context enhancement and density prior," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7040–7051, 2024.
[29] W. Yang, S. Vitale, H. Aghababaei, G. Ferraioli, V. Pascazio, and G. Schirinzi, "A deep learning solution for height estimation on a forested area based on Pol-TomoSAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5208214.
[30] M. Schwartz et al., "High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach," *Int. J. Appl. Earth Observ.*, vol. 128, 2024, Art. no. 103711.
[31] L. Li et al., "Ultrahigh-resolution boreal forest canopy mapping: Combining UAV imagery and photogrammetric point clouds in a deep-learning-based approach," *Int. J. Appl. Earth Observ.*, vol. 107, 2022, Art. no. 102686.
[32] G. Morales, G. Kemper, G. Sevillano, D. Arteaga, I. Ortega, and J. Telles, "Automatic segmentation of mauritia flexuosa in unmanned aerial vehicle (UAV) imagery using deep learning," *Forests*, vol. 9, no. 12, 2018, Art. no. 736.
[33] Z. Ye et al., "Extraction of olive crown based on UAV visible images and the U2-net deep learning model," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1523.
[34] I. Jamaluddin, Y.-N. Chen, and K.-C. Fan, "Spatial–spectral–temporal deep regression model with convolutional long short-term memory and transformer for the large-area mapping of mangrove canopy height by using Sentinel-1 and Sentinel-2 data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4403117.
[35] G. Li, W. Han, S. Huang, W. Ma, Q. Ma, and X. Cui, "Extraction of sunflower lodging information based on UAV multi-spectral remote sensing and deep learning," *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2721.
[36] Y. Xie et al., "Instance segmentation and stand-scale forest mapping based on UAV images derived RGB and CHM," *Comput. Electron. Agriculture*, vol. 220, 2024, Art. no. 108878.
[37] G. Modica, G. Messina, G. De Luca, V. Fiozzo, and S. Praticò, "Monitoring the vegetation vigor in heterogeneous citrus and olive orchards. A multiscale object-based approach to extract trees' crowns from UAV multispectral imagery," *Comput. Electron. Agriculture*, vol. 175, 2020, Art. no. 105500.
[38] P. Martínez-Carricondo, F. Agüera-Vega, F. Carvajal-Ramírez, F.-J. Mesas-Carrascosa, A. García-Ferrer, and F.-J. Pérez-Porras, "Assessment of UAV-photogrammetric mapping accuracy based on variation of ground control points," *Int. J. Appl. Earth Observ.*, vol. 72, pp. 1–10, 2018.
[39] H. He, J. Zhou, M. Chen, T. Chen, D. Li, and P. Cheng, "Building extraction from UAV images jointly using 6D-SLIC and multiscale Siamese convolutional networks," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1040.
[40] D. Booth, S. E. Cox, T. Meikle, and C. Fitzgerald, "The accuracy of ground-cover measurements," *Rangeland Ecol. Manage.*, vol. 59, no. 2, pp. 179–188, 2006.
[41] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.*, vol. 8, no. 2, pp. 127–150, 1979.
[42] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.
[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

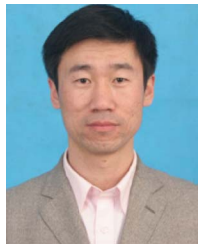
[44] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.
[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, vol. 9351, pp. 234–241.
[46] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, Dec. 2020, Art. no. 71.
[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoderdecoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[48] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[49] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[50] X. Dong et al., "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12124–12134.
[51] S. Woo et al., "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.
[52] Q. Zhu et al., "Samba: Semantic segmentation of remotely sensed images with state space model," 2024, *arXiv:2404.01705*.
[53] M. Fromm, M. Schubert, G. Castilla, J. Linke, and G. McDermid, "Automated detection of conifer seedlings in drone imagery using convolutional neural networks," *Remote Sens.*, vol. 11, no. 21, Nov. 2019, Art. no. 2585.
[54] J. R. G. Braga et al., "Tree crown delineation algorithm based on a convolutional neural network," *Remote Sens.*, vol. 12, no. 8, Apr. 2020, Art. no. 1288.

Haiqing He received the Ph.D. degree in geodesy and surveying engineering from Wuhan University, Wuhan, China, in 2013.

From 2017 to 2019, he was with the Lyles School of Civil Engineering, Purdue University, IN, USA, as a Post-Doctoral Fellow. He is currently a Full Professor with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China, and also with Jiangxi Key Laboratory of Watershed Ecological Process and Information, East China University of Technology, Nanchang, China. His research interests include low-attitude photogrammetry, image matching, artificial intelligence driven remote-sensing image interpretation, and 3D reconstruction.

Fuyang Zhou received the B.S. degree in surveying and mapping engineering, in 2021, from East China University of Technology, Nanchang, China, where he is currently working toward the Ph.D. degree in surveying and mapping.

His current research interests include photogrammetry and remote sensing, image processing, and machine learning.



Yongjun Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in geodesy and surveying engineering from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist with Environmental Systems Research Institute Inc. (Esri), Redlands, CA, USA. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. He holds 25 Chinese patents and 26 copyrights registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence driven remote-sensing image interpretation, integration of LiDAR point clouds and images, and 3-D city reconstruction.

Dr. Zhang is the PI Winner of the Second-Class National Science and Technology Progress Award in 2017 and the PI Winner of the Outstanding-Class Science and Technology Progress Award in Surveying and Mapping (Chinese Society of Surveying, Mapping and Geoinformation, China) in 2015. He is a Key Member of ISPRS Workgroup II/I from 2016 to 2020. He is the Coeditor-in-Chief of *The Photogrammetric Record*.

Ting Chen received the B.S. and Ph.D. degrees in water conservancy engineering from Wuhan University, Wuhan, China, in 2013 and 2016, respectively.

She is currently a Lecturer with the School of Water Resources and Environmental Engineering, East China University of Technology, Nanchang, China. Her research interests include photogrammetry and remote sensing.

Yan Wei received the B.S. degree in surveying and mapping engineering from Nanyang Normal University, Nanyang, China, in 2021. She is currently working toward the M.S. degree in surveying and mapping engineering with East China University of Technology, Nanchang, China.

Her current research interests include photogrammetry and remote sensing, image processing, and machine learning.