Parallel Fusion Neural Network Considering Local and Global Semantic Information for Citrus Tree Canopy Segmentation

Haiqing He, Fuyang Zhou, Yuanping Xia, Min Chen, and Ting Chen

***Abstract*—Existing convolutional neural network (CNN) based methods usually tend to ignore the contextual information for citrus tree canopy segmentation. Although popular transformer models are helpful in extracting global semantic information, they ignore the edge details between citrus tree canopies and the background. To address these issues, we propose a parallel fusion neural network considering both local and global semantic information for citrus tree canopy segmentation from 3-D data, which are derived by unmanned aerial vehicle (UAV) mapping. In the feature extraction stage, a parallel architecture, concatenated by EfficientNet-V2 and CSwin transformer, is used to extract local and global information of citrus trees. In the feature fusion stage, we design a coordinate attention-based fusion module to retain the contextual information and local edge details of citrus tree canopies. Additionally, to exaggerate the exclusivity between tree canopies and complex backgrounds, 3-D data incorporating RGB imagery and canopy height model derived by UAV photogrammetry are generated for citrus tree canopy segmentation. Experimental results indicate that the proposed method performs considerably better than methods based only on CNN or transformer models and is superior to state-of-the-art methods (e.g., the highest mIoU score of 93.46%).**

***Index Terms*—Citrus tree canopy, complex background, contextual information, self-attention mechanism, semantic segmentation.**

Manuscript received 8 July 2023; revised 29 September 2023; accepted 24 October 2023. Date of publication 5 December 2023; date of current version 15 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42261075, Grant 41861062, Grant 42174055, and Grant 42074005, in part by Jiangxi Provincial Natural Science Foundation under Grant 20224ACB212003, 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, and in part by the Innovation Fund Designated for Graduate Students of East China University of Technology under Grant DHYC-202327. *(Corresponding author: Haiqing He.)*

Haiqing He, Fuyang Zhou, and Yuanping Xia are with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China, and also with the Key Laboratory of Mine Environmental Monitoring and Improving Around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China (e-mail: hyhqing@163.com; fuy_zhou@163.com; ypxia@ecut.edu.cn).

Min Chen is with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: minchen@home.swjtu.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).

Digital Object Identifier 10.1109/JSTARS.2023.3339290

I. Introduction

CITRUS tree positioning and counting are conducive to high-throughput phenotypic research and fine agricultural management. The most important component of a citrus tree is the canopy, which is usually considered an indicator for evaluating growth vitality and characterizing the competitive relationship between trees [1]. Therefore, obtaining canopy information is crucial to determining the location, quantity, and growth of citrus trees. Traditionally, manual surveys and field measurements of citrus tree canopies are time-consuming and labor-intensive, which are often unable to satisfy the requirement of high-efficiency and high-accuracy acquisition of citrus tree canopy information. In this study, to obtain citrus tree canopy information efficiently and low costly, we concentrate on how to extract citrus tree canopy information from ultrahigh-resolution and low-cost unmanned aerial vehicle (UAV) photogrammetry-derived data. In terms of image-based information extraction, image segmentation is the crucial step for tree canopy extraction.

In recent years, given the great success of deep learning, semantic segmentation methods based on deep learning have attracted continuous and even increasing attention from researchers [2], [3]. Compared with manually designed methods, deep learning-based segmentation methods perform considerably better in many applications [4], [5], [6]. The fully convolutional network (FCN) [7], which consists of convolutional and pooling layers arranged alternately, is the first end-to-end semantic segmentation neural network. To alleviate the loss of local details caused by pooling operations in FCNs, U-Net [8] was proposed to retain the local details of the original input size of the image by connecting low-level and high-level features through skip connections' operations. However, given the limited receptive field of the convolutional kernel in FCN or U-Net, it cannot extract rich the global contextual information of the image. To address this problem, some improved deep networks were proposed. Chen et al. [9], [10] presented DeepLab to increase the receptive field of the convolutional kernel through dilated convolutions. In addition, other networks, such as DANet [11], the bottleneck attention module [12], and the convolutional block attention module [13], have been proposed to extract global semantic information by introducing attention mechanisms. By increasing the receptive field or introducing an attention mechanism, the problem of insufficient global semantic information for image segmentation can be alleviated to varying degrees, but

the global information obtained by these two methods is limited. In contrast to the convolutional neural network (CNN) used in the aforementioned deep networks, the transformer model proposed by Google based on an attention mechanism can effectively obtain global contextual information [14]. Derived deep networks based on the transformer model have been proposed for image applications. For example, the vision transformer (ViT) [15] model was proposed for image processing, which achieved excellent performance on large-scale datasets due to the use of global contextual information. In addition, a model specifically proposed for image segmentation, named segmentation transformer (SETR) [16], was applied to conduct semantic segmentation tasks by introducing CNN as a decoder based on ViT. Since then, many researchers have proposed a series of improved algorithms based on the transformer model, such as ResT [17], BEiT [18], Swin [19], and CSwin [20]. Given the outstanding capabilities of local and global feature extraction by using CNN and transformer, these semantic segmentation models have been proven to perform well in object segmentation [21], [22].

In addition, inspired by the powerful performance of land cover semantic segmentation from remote sensing images [23], [24], [25], a state-of-the-art deep network considering local and global semantic information, namely, DeepLab V3+ [10], was used for tree canopy segmentation in different scenarios [26]. Guirado et al. [27] and Braga et al. [28] used the mask R-CNN instance segmentation model to segment tree canopies in tropical forests and drylands. Except for the visible-spectrum images used in the above methods, multiband images, including the near-infrared band, were also used for tree canopy segmentation in deep networks and achieved better performance compared with using only the visible-light bands [29], [30]. Specifically, related to this study, deep neural networks have been widely studied for tree canopy segmentation. Different CNN-based models, such as FCN, U-Net, SegNet, DeepLabV3+, DDCN, SSD, R-CNN, and faster R-CNN, were applied to extract tree canopies [31], [32], [33]. However, these deep networks have difficulty satisfying the requirements of citrus tree canopy segmentation under complex terrain and backgrounds, mainly due to the following two reasons: First, tree canopy segmentation relying solely on 2-D images is insufficient to characterize the uniqueness of tree canopies; and second, most CNN-based methods that do not consider global contextual information are sensitive to local information around tree canopies. Therefore, how to effectively integrate the advantages of CNN and transformer in extracting local and global contextual information of tree canopies, and introduce extra information to improve the applicability of deep networks, has become a crucial and valuable issue for tree canopy segmentation.

According to the above-mentioned literature, CNNs can effectively extract local detailed information but lack global contextual information. By contrast, the transformer can effectively extract global contextual information for each pixel but lacks local detail information. Hence, the combination of CNN and transformer through transfer learning can help improve the performance of citrus tree canopy segmentation under complex backgrounds. In addition, the height and geometric structure of citrus tree canopies, characterized by the canopy height model (CHM), are valuable to eliminate the influence of terrain and broaden the discriminative gap between citrus trees and surrounding weeds. Therefore, in this study, a parallel fusion neural network considering local and global semantic information is proposed to cater to the requirements of citrus tree canopy segmentation under complex backgrounds. This network consists of two branches, including CNN and transformer channels. The CNN branch mainly extracts local features between adjacent pixels, whereas the transformer branch mainly extracts global contextual information between pixels in the entire image. To combine the advantages of the two branches effectively, we design a coordinate attention-based fusion module (CAFM) to retain the contextual information and local edge details of citrus tree canopies. Considering the influence of terrain and weeds fully, the CHM without topographic relief is input into the proposed network along with the visible-spectrum image. That is, the input data can be considered 3-D (i.e., 2-D true-color RGB image and CHM).

The main contributions are given as follows.

1) A parallel deep network combining improved EfficientNet-V2 and CSwin transformer was designed to extract local and global semantic information, by which the contextual information and local edge details of citrus tree canopies can be effectively retained.
2) In the decoding phase of the proposed network, a feature fusion module was constructed to integrate local and global semantic information extracted by EfficientNet-V2 and CSwin transformer, respectively. It is very useful for reducing redundant information that is not valuable for citrus tree canopy segmentation.
3) The 3-D data used in this study combines RGB imagery and CHM derived by UAV photogrammetry, which can exaggerate the exclusivity between tree canopies and complex backgrounds (terrain variations and dense vegetation especially).
4) By sharing the initial weights and parameters of state-of-the-art deep networks through transfer learning, it is possible to train the proposed network without requiring a large amount of training sample data from scratch.

II. Related Works

This section mainly introduces the research progress of semantic segmentation methods based on deep learning, including CNN-based semantic segmentation methods, transformer-based semantic segmentation methods, and CNN- and transformer-coupled segmentation methods, as well as some related work on semantic segmentation in tree canopy segmentation.

A. Semantic Segmentation Methods Based on Deep Learning

As the first end-to-end semantic segmentation fully convolutional network, i.e., FCN [7], it promotes the application of deep learning models in image segmentation. However, the pooling operation in FCN destroys the spatial information of feature maps (such as shape and texture), resulting in a lack of local detailed information in FCN. To obtain high-level semantic

representation while retaining the local detailed information of objects, U-Net [8] improved the segmentation ability of CNN in local details effectively by fusing high-resolution feature maps at different levels in the encoding phase with skip connections. Given the limited receptive field of convolutional kernels, CNN-based semantic segmentation models also lack global contextual information. To obtain global contextual information, some researchers have proposed different multiscale contextual information fusion methods. DeepLab was proposed to increase the receptive field of convolutional kernels effectively by extracting multiscale contextual information using an atrous spatial pyramid pooling module with a dilated convolution operation [9], [10]. PSPNet was proposed to fuse global contextual information with different receptive fields by using the pyramid pooling module [34]. In addition, some researchers have introduced attention mechanisms [14] to establish global relationships. DANet [11] was designed to establish long-term dependency relationships and obtain global contextual information from different dimensions by a dual attention mechanism for channel and spatial dimensions, and HRCNet [35] was designed to obtain global information by a lightweight dual attention module. The above-mentioned networks can all obtain global information through multiscale information fusion and attention mechanisms, but their ability to extract global contextual information remains limited.

Compared with CNN-based methods, the transformer methods based on self-attention mechanisms can effectively obtain global contextual information, thus showing excellent performance in the field of computer vision [3]. The aforementioned transformer models (e.g., ViT [15] and SETR [16]) require a series of patches for input, thus ignoring the local information in each patch. To obtain local representations in the transformer structure, many researchers have proposed improved algorithms based on the transformer model. For example, Zhang and Yang [17] designed a deep network, namely, ResT, to obtain the feature maps of different levels and semantic representations by a patch embedding layer. Liu et al. [19] designed a deep network named Swin, which divides the input image into nonoverlapping small images of size 4×4 and performs self-attention calculation in each small image. Subsequently, based on Swin and local enhanced position encoding (LePE), CSwin [20] was proposed to obtain local information by performing self-attention calculations within the cross-shaped stripe window. Although the above transformer-based methods can obtain the local information of images, the transformer-based pure attention mechanism is weaker than CNN in obtaining local detailed information. How to obtain local and global information simultaneously by reasonably fusing CNN and transformer has become an important issue in semantic segmentation.

Therefore, to combine the advantages of CNN and transformer, some researchers have proposed different fusion methods. For example, Chen et al. [36] proposed TransUNet in which skip connections are introduced to fuse shallow features extracted by CNN, and global contextual information extracted by transformer is combined to segment together. Zhang et al. [37] used Swin transformer and CNN as the encoder and decoder, respectively, and adopted the spatial–asymmetric spatial pyramid pooling module based on deep separable convolution to obtain multiscale contextual information. In addition to the aforementioned serial fusion methods, the parallel fusion methods can also be used to fuse the CNN and transformer. Zhang et al. [38] proposed a parallel network architecture named TransFuse in which the BiFusion module is used to fuse the multilevel features, including global contextual information and local detail information, which are extracted by CNN and transformer, respectively. Subsequently, Gao et al. [39] proposed an improved TransFuse called STransFuse, which combines ResNet and Swin transformer in a parallel structure and fuses global and local information through an adaptive fusion module, achieving good performance on the Vaihingen and Potsdam remote sensing datasets. Although the aforementioned CNN and transformer fusion methods can achieve good performance in medical images and public datasets, they are not suitable for land cover segmentation in complex backgrounds. The main reason is that these methods are easily influenced by ground objects with similar textures, thus leading to poor segmentation accuracy.

B. Tree Canopy Segmentation Based on Deep Learning

Closely related to this study, many other studies have also been conducted on tree canopy semantic segmentation based on deep learning from remote sensing imagery. Typically, Morales et al. [26] used the DeepLabV3+ network to segment palm tree canopies from UAV high-resolution imagery, achieving an accuracy of 98.14% on the test set. Braga et al. [28] used the mask R-CNN network to detect and delineate tree canopies in tropical forests, obtaining good tree crown segmentation results in high-resolution satellite imagery. Guirado et al. [27] fused mask R-CNN with object-based image analysis to segment the scattered vegetation in the arid ecosystem, demonstrating a 25% higher accuracy compared with a single model. These methods can achieve high segmentation accuracy in specific environments but do not consider using multisource data to further improve the accuracy of tree canopy segmentation. Li et al. [29] input multiband imagery into networks, such as SegNet and U-Net, to extract the semantic information of large-area sunflower lodging and achieved an accuracy of 88.23%. Hao et al. [30] input six bands of UAV remotely sensed imagery into the mask R-CNN network to segment tree canopies and inferred that additional bands could remarkably improve the performance of tree canopy segmentation.

In addition, different semantic models based on CNNs have been applied to evaluate the performance of tree canopy segmentation. Fromm et al. [31] compared three CNN networks, i.e., SSD, R-CNN, and faster R-CNN, to segment seedlings in large areas of coniferous forests automatically. Martins et al. [32] used five CNN networks, i.e., FCN, U-Net, SegNet, DDCN, and DeepLabV3+, to evaluate the performance of tree canopy segmentation in urban environments from true-color aerial RGB imagery, achieving an average accuracy of 91.25%. However, these CNN-based methods perform tree canopy segmentation under simple background conditions, such as flat terrain, without considering factors, such as terrain fluctuations, spectral similarity of weeds, and complex spatial information. These methods

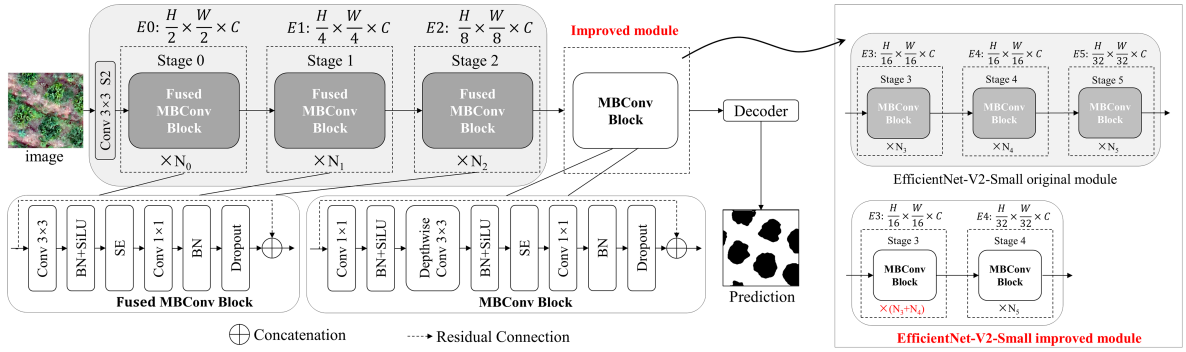

Fig. 1. Architecture of the improved EfficientNet-V2.

are mainly based on 2-D image segmentation without considering 3-D information, such as the height of ground objects, which may easily lead to misclassification of tree canopy pixels. Thus, the performance of tree canopy segmentation is affected by local information and global contextual information; the former is strongly related to tree canopy delineation, and the latter is crucial to separate tree canopies from image backgrounds. However, the current tree canopy segmentation methods are all based on CNNs, and studies on the fusion of CNN and transformer for tree canopy segmentation are few. Therefore, it is of great significance to explore a multisource data-based semantic segmentation method to satisfy the requirements of citrus tree canopy segmentation in the case of terrain undulations and complex backgrounds by extracting both local and global information.

III. Methods

This section introduces a lightweight CNN model (i.e., EfficientNetV2 [40]) and the latest transformer model (i.e., CSwin [20]). On the basis of the advantages of these models, a fusion neural network of CNN and transformer is proposed. Finally, the loss function used in this study is introduced in detail.

A. Transfer Learning Based on EfficientNet-V2 and CSwin Transformer

Inspired by the progress of EfficientNet-V2 and CSwin transformer in local and global semantic information extraction [20], [40], this study introduces the two networks for citrus tree segmentation through transfer learning.

Noticeably, as illustrated in the block "Improved module" in Fig. 1, the structure of EfficientNet-V2 is improved by merging the first two MBConv stages in the EfficientNet-V2-S scale into one MBConv stage and merging the last four MBConv stages in the EfficientNet-V2-M and EfficientNet-V2-L scales into two MBConv stages so that it enables the output features of the last four stages of the network to be fused with CSwin transformer. The overall architecture of the improved EfficientNet-V2, which consists of five stages based on the fused MBConv and MBConv modules, is illustrated in Fig. 1. The outputs of each stage are $\frac{H}{2} \times \frac{W}{2}$, $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$. By adjusting the number of MBConv and fused MBConv modules in each stage ($\times$Ni), three different scale sizes of EfficientNet-V2 can be formed, namely, EfficientNet-V2-S, EfficientNet-V2-M, and EfficientNet-V2-L. Through the hierarchical architecture of the improved EfficientNet-V2, multiscale semantic information can be obtained and well adapted to downstream segmentation tasks. Therefore, the local features extracted by the improved EfficientNet-V2 can supplement the insufficient local information extracted in the transformer model.

In this study, CSwin transformer is used to capture global semantic information. In contrast to the CNN network, in the CSwin transformer, the input patches with size of $H \times W \times 3$ are downsampled into size $\frac{H}{4} \times \frac{W}{4}$ and transformed into the channel dimension of C by an overlapped convolutional token embedding layer (7×7 convolutional layer with stride 4). To obtain multiscale global semantic information, a hierarchical representation is designed. CSwin transformer consists of four stages in which the dimensions of the output feature maps are $\frac{H}{4} \times \frac{W}{4} \times C$, $\frac{H}{8} \times \frac{W}{8} \times 2C$, $\frac{H}{16} \times \frac{W}{16} \times 4C$, and $\frac{H}{32} \times \frac{W}{32} \times 8C$. By adjusting the number of CSwin transformer modules in each stage ($\times$Ni), four different scale sizes of CSwin transformer can be formed, namely, CSwin-Tiny, CSwin-Small, CSwin-Base, and CSwin-Large. The structure of CSwin transformer is similar to that of ResNet [41], which extracts different scale information of images through a hierarchical representation to adapt to pixel-level semantic segmentation tasks and uses residual connections to avoid gradient vanishing. In addition, an attention mechanism named cross-shaped window self-attention was designed in the CSwin transformer module to calculate self-attention in the horizontal and vertical stripes of the cross-shaped window parallelly. Compared with the commonly used full attention mechanisms, this design can effectively reduce the computational cost of the network. In addition, LePE was introduced in the CSwin transformer module to enhance local information further. Hence, the powerful performance of CSwin transformer in extracting global context information can be well applied to global information extraction in the semantic segmentation task of this study.

B. Fusion Neural Network Considering Local and Global Semantic Information

Based on the advantages of EfficientNet-V2 and CSwin transformer, in this study, a parallel fusion neural network considering

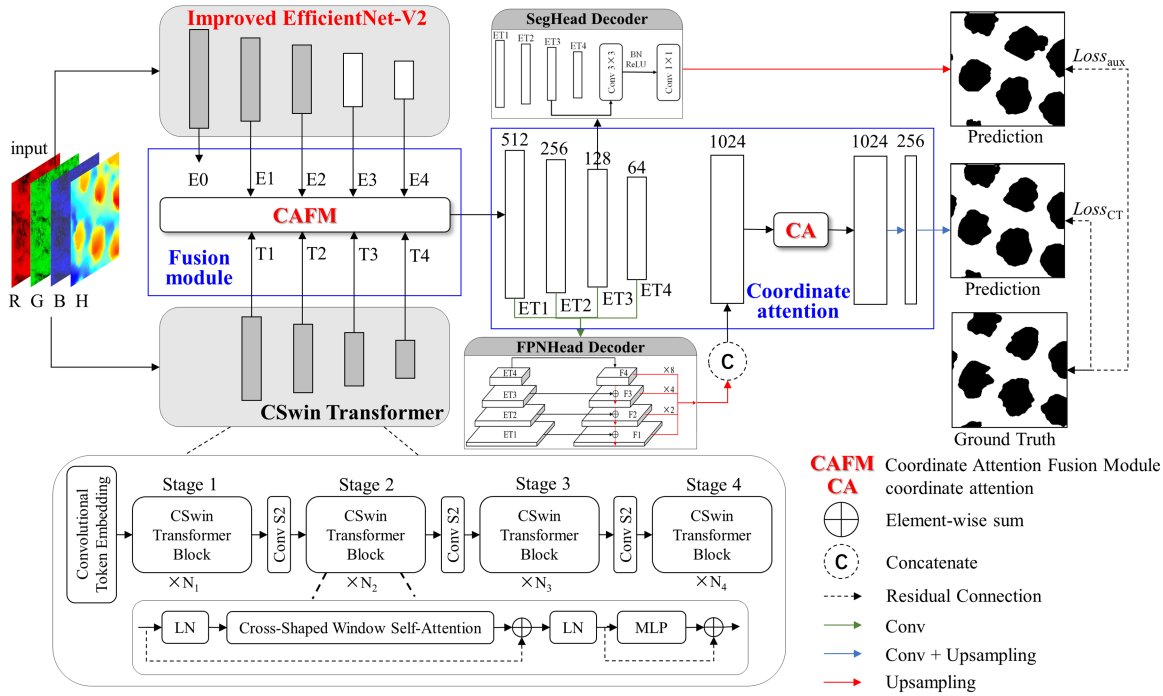

Fig. 2. Architecture of the proposed fusion neural network.

local and global semantic information for citrus tree canopy segmentation is proposed, and its architecture is shown in Fig. 2.

The architecture of the proposed fusion neural network consists of four parts:

1) local semantic information extraction using improved EfficientNet-V2;
2) global semantic information extraction using CSwin transformer;
3) CAFM;
4) decoder module.

First, the improved EfficientNet-V2 can generate five features $E0$, $E1$, $E2$, $E3$, and $E4$, with resolutions of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input resolution, respectively. CSwin transformer parallelly generates four features $T1$, $T2$, $T3$, and $T4$, with resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input resolution, respectively. Second, the proposed CAFM fusion module is explored to fuse the features of $E1$ and $T1$, $E2$ and $T2$, $E3$ and $T3$, and $E4$ and $T4$ to obtain feature maps with local and global contextual information. Here, the CAFM fusion module is a feature reconstruction module based on the coordinate attention mechanism (briefly called CA) [42], which not only retains the extracted local and global information as much as possible but also preserves object positional information in the feature information. Subsequently, the fused features $ET1$, $ET2$, $ET3$, and $ET4$, which are obtained by the proposed CAFM, are input into the FPNHead decoder module to generate four types of feature maps with the same size and channel number by multiscale feature fusion. The four types of feature maps are concatenated into a 1024-channel feature map, which is then input into the CA module to extract the object's positional information further. Finally, the output features of the CA module are input into the convolutional and upsampling layers, and upsampled to the original input image size to obtain a complete semantic segmentation result.

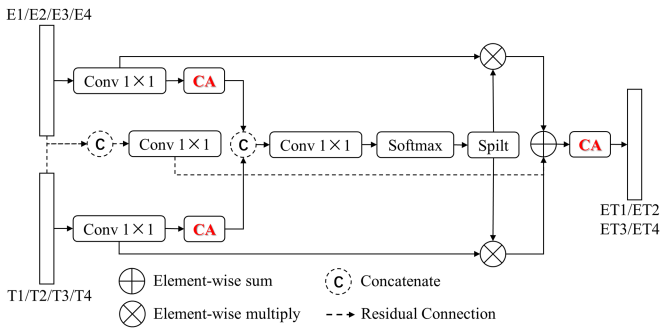

Fig. 3. Architecture of the proposed CAFM.

C. CA-Based Fusion Module

To achieve the purpose of combining local and global semantic information, we design a feature fusion module (i.e., CAFM) based on coordinate attention (CA) mechanism [42], as shown in Fig. 3. This module adaptively fuses semantic information between features of different scales by using self-attention mechanisms. The 2-D pooling operation in the conventional attention mechanism can easily lead to the loss of spatial information and high computational costs. In contrast, the CA mechanism utilizes two lightweight 1-D pooling operations to aggregate horizontal and vertical spatial perception, which enables the CA to accurately locate citrus trees. Therefore, to effectively locate and distinguish the boundaries between citrus trees and backgrounds (such as shrubs and weeds), the CA was introduced into

the feature fusion module (CAFM) to capture spatial position information.

First, the features (i.e., *E*1, *E*2, *E*3, *E*4, *T*1, *T*2, *T*3, and *T*4 called *ET*) obtained by EfficientNet-V2 and CSwin transformer are input to a 1×1 convolutional transformation function F_1 to yield a specific number of channels. Second, each feature map, such as y_E, y_T, and y_{ET}, is extracted with the object's positional information through a CA [42]. The generation of y_E, y_T, and y_{ET} can be mathematically expressed as

$$y_i = \begin{cases} \mathrm{CA}(F_1(X_i)), & i \in (E, T) \\ F_1([Xi, Xj]), & i \in E, j \in T \end{cases} \tag{1}$$

where $\mathrm{CA}(\cdot)$ denotes the coordination attention operation, F_1 denotes the convolutional transformation function, X denotes the variable for an input feature, and $[\cdot, \cdot]$ denotes the concatenation operation along the spatial dimension. Third, the feature maps calculated by CA are concatenated through concatenation operations along the spatial dimension, and the information interaction between the improved EfficientNet-V2 and CSwin transformer is achieved through a 1×1 convolutional operation. In addition, to accelerate the training convergence of CAFM, one residual connection operation is used at the beginning of the two branches of features *E* and *T*. Fourth, the SoftMax function is used to calculate the weight of each pixel in the feature map. Then, a split operation $\mathrm{split}(\cdot)$ is explored to separate the weight map into two feature maps y'_E and y'_T, which are more capable of characterizing objects of interest

$$y'_E, y'_T = \mathrm{split}(\delta(F_1([y_E, y_T]))) \tag{2}$$

where δ is a nonlinear activation function (i.e., SoftMax in this study). Finally, the two feature maps y'_E and y'_T are concatenated and input into a CA module to extract the object positional information in the fused feature map X' of the CNN and transformer. The mathematical module can be expressed as

$$X' = \mathrm{CA}(\mathrm{sum}(F_1(X_E) \cdot y'_E, F_1(X_T) \cdot y'_T, y_{ET})) \tag{3}$$

where $\cdot$ denotes the dot multiplication, and $\mathrm{sum}(\cdot)$ is the elementwise sum operation.

As shown in Fig. 4, several typical feature maps from *E*2, *T*2, and *ET*2 obtained by CNN, transformer, and CAFM, respectively, are selected to demonstrate the effectiveness of CAFM. Compared with the feature maps obtained by CNN, it can be seen that CNN usually has higher weight on edge details, which enables citrus canopy boundaries to be more discriminative. However, CNN-based methods are limited by their receptive field and usually tend to ignore contextual information, making it difficult to distinguish shrubs with similar spatial information to citrus tree canopies. Different from CNN, the transformer has the ability of long-distance dependencies and can extract global semantic information, which is helpful to accurately distinguish tree canopies from backgrounds. However, as shown in the third column of Fig. 4, due to the lack of local details, the feature maps obtained by the transformer focus more on backgrounds with higher weight rather than citrus tree canopies, resulting in the transformer being unable to accurately extract local details, such as citrus tree canopy boundaries. To make full use of the advantages of CNN and transformer, CAFM is proposed to characterize and segment citrus tree canopies. In contrast, as shown in the fourth column of Fig. 4, the proposed network considering local and global semantic information performs significantly better than CNN and transformer, as it can exaggerate the exclusivity between canopies and backgrounds.

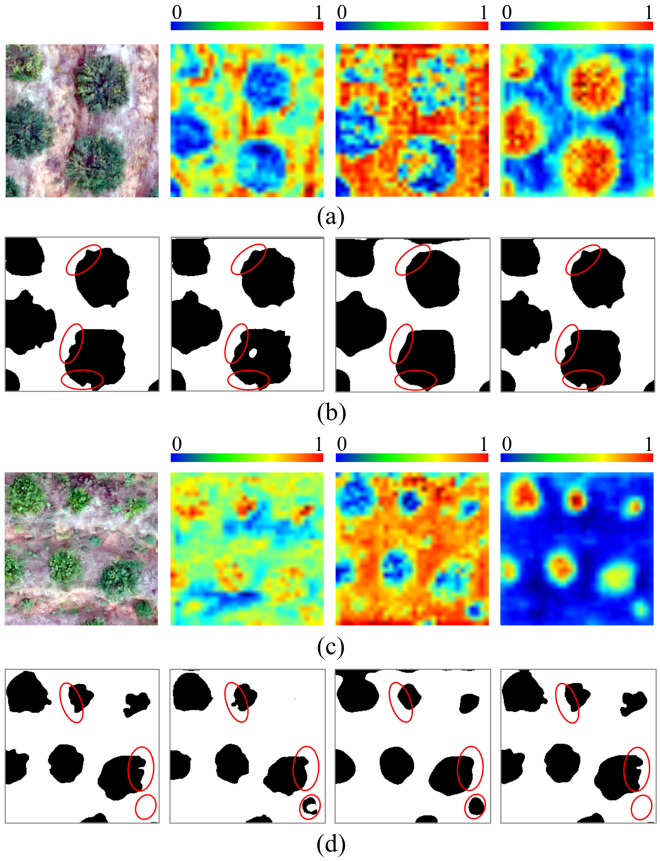

Fig. 4. Comparison of feature maps and segmentation results obtained by CNN, transformer, and CAFM. The first to fourth columns of (a) and (c) are UAV images, feature maps obtained by CNN, transformer, and CAFM, respectively. The first to fourth columns of (b) and (d) represent the ground truth, segmentation results obtained by CNN, transformer, and CAFM, respectively.

The main loss function $\mathrm{Loss}_{\mathrm{CT}}$ is used to supervise the training from the improved EfficientNet-V2 and CSwin transformer to CA. In addition, in the proposed fusion neural network, an auxiliary loss $\mathrm{Loss}_{\mathrm{aux}}$ is used to supervise branches (for generating one of the fused feature *ET*3) to improve the convergence performance during error backpropagation.

D. Loss Function

The loss function is used to calculate the error between the predicted value and the ground-truth value during network training, which is crucial for optimizing model parameters. Generally, the cross-entropy (CE) function is the commonly used loss function in the field of semantic segmentation. Given the varying sizes of tree canopies, the area of the tree canopy relative to the background area is relatively small, resulting in an imbalance between foreground and background in the samples. Therefore, in this study, to alleviate the problem of network bias toward the background nonvariable classes, the dice coefficient loss

function [43] is introduced to orient attention toward the citrus tree instead of the background. Meanwhile, to obtain the loss of each pixel in the input image $\mathbb{R}^{H\times W}$, the loss of $H \times W$ pixels is summed and averaged.

Specifically, the total loss function $\text{Loss}_{\text{total}}$ used in this study is a mixed loss function of the binary CE function and dice coefficient loss function, and the expressions of the two loss functions are given in (4) and (5). The mixed loss function (i.e., $\text{Loss}_{\text{total}}$) consists of a primary loss function (i.e., Loss_{CT}) and an auxiliary loss function (Loss_{aux}), expressed as (6). Loss_{CT} is used to supervise the optimization process of the entire network in the encoding and decoding phases, whereas Loss_{aux} is used to supervise the learning of features *ET*3 and enhance the characteristics of the fused features. The two loss functions Loss_{CT} and Loss_{aux} can be calculated by (7). Noticeably, to optimize the proposed module from different loss calculation perspectives in different training samples, a factor α is given to adjust the contribution of Loss_{CT} and Loss_{aux}, and α is set to 0.5 determined by achieving the best performance of this study based on multiple tries. The loss functions $\text{Loss}_{\text{total}}$, Loss_{CT}, and Loss_{aux} are mathematically expressed as

$$\text{BCELoss} = -\frac{1}{H\times W}\sum_{i=1}^{H\times W}\sum_{n=1}^{N} y_{i,n}\log(\hat{y}_{i,n}) \tag{4}$$

$$\text{DiceLoss} = \frac{1}{H\times W}\left(1-\frac{2\sum_{i=1}^{H\times W} y_i\hat{y}_i}{\sum_{i=1}^{H\times W} y_i+\sum_{i=1}^{H\times W}\hat{y}_i}\right) \tag{5}$$

$$\text{Loss}_{\text{total}} = \text{Loss}_{\text{CT}}+\text{Loss}_{\text{aux}} \tag{6}$$

$$\text{Loss}_{\text{CT}}(\text{Loss}_{\text{aux}}) = \alpha\times\text{BCELoss}+(1-\alpha)\times\text{DiceLoss} \tag{7}$$

where y_i represents the *i*th pixel in the ground-truth values, $\hat{y}_i$ represents the predicted value corresponding to y_i, *n* represents the *n*th class, *N* is the number of classes, and $y_{i,n}$ represents a symbolic function that equals 1 if the *i*th pixel belongs to the *n*th class and 0 otherwise. $\hat{y}_{i,n}$ represents the probability that the *i*th pixel is predicted as the *n*th class.

E. Evaluation Metric

This study uses four evaluation metrics, including overall accuracy (OA), precision, recall, *F*1 score, and intersection over union (IoU), to evaluate the performance of tree crown segmentation. The formulas for calculating these metrics are as follows:

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{10}$$

$$F1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{12}$$

where TP represents the number of correctly predicted canopy pixels, FP represents the number of incorrectly predicted pixels that are classified as canopy while they belong to the background, TN represents the number of correctly predicted background pixels, and FN represents the number of incorrectly predicted pixels that belong to the canopy while they are classified as background.

IV. Experiments

A. Study Area

To verify the effectiveness of the proposed method for segmenting citrus trees in complex backgrounds, the study area (located in Xinfeng County, Jiangxi Province, China) with large terrain variations and surrounding dense vegetation was specifically selected. Given the acidic soil conditions and the typical subtropical monsoon humid climate with abundant rainfall and sufficient sunlight, the study area is suitable for planting citrus trees, such as navel orange trees. To compare and analyze the applicability of the proposed method under different terrain and canopy background conditions, we extract four subareas of the study area in Fig. 5; the areas are described in detail as follows.

1) The terrain of Plots 1 and 2 is highly undulating, with an altitude of 135–177 m, and no adhesion occurs between the citrus tree canopies. Several subregions of Plots 1 and 2 include shrubs and weeds, and their visible spectra are similar to those of citrus trees. In contrast to Plot 1, shrubs or weeds in Plot 2 have a considerable gap that exists between shrubs or weeds and citrus tree canopies, and the canopy spacing is relatively large.
2) The terrain of Plot 3 is relatively flat, but the canopies are highly adhesive; moreover, many shrubs and weeds have spectra similar to those of citrus tree canopies.
3) Plot 4 combines the characteristics of Plot 1, Plot 2, and Plot 3, with highly undulating, dense shrubs and weeds, and adhesive tree canopies. The four subareas planted only one type of fruit tree (i.e., citrus tree), and citrus trees of different heights and sizes were widely distributed in Plots 1 and 4 because of different growth periods caused by replanting.

B. Data Processing

In this study, high-resolution overlapping images were collected by a small quadcopter UAV (DJI Phantom 4 RTK, DJI, Shenzhen, China) for aerial triangulation to generate a digital orthophoto model (DOM) and digital surface model (DSM) of the study area. The UAV image acquisitions were performed from October 9–12, 2022 under good weather conditions, such as sunny and winds of $<$10 m/s. The flight speed was 5 m/s, and the relative flight altitude of the UAV was approximately 80 m, which led to capturing UAV remote sensing images with a spatial resolution of approximately 3 cm/pix. The overlap of aerial stereo images was set to 80% to ensure sufficient overlaps in the case of large terrain fluctuations in the study area. A total of 856, 794, 672, and 812 images were captured for Plot 1, Plot 2, Plot 3, and Plot 4, respectively, with a frame size of 5472 $\times$ 3648 pixels.

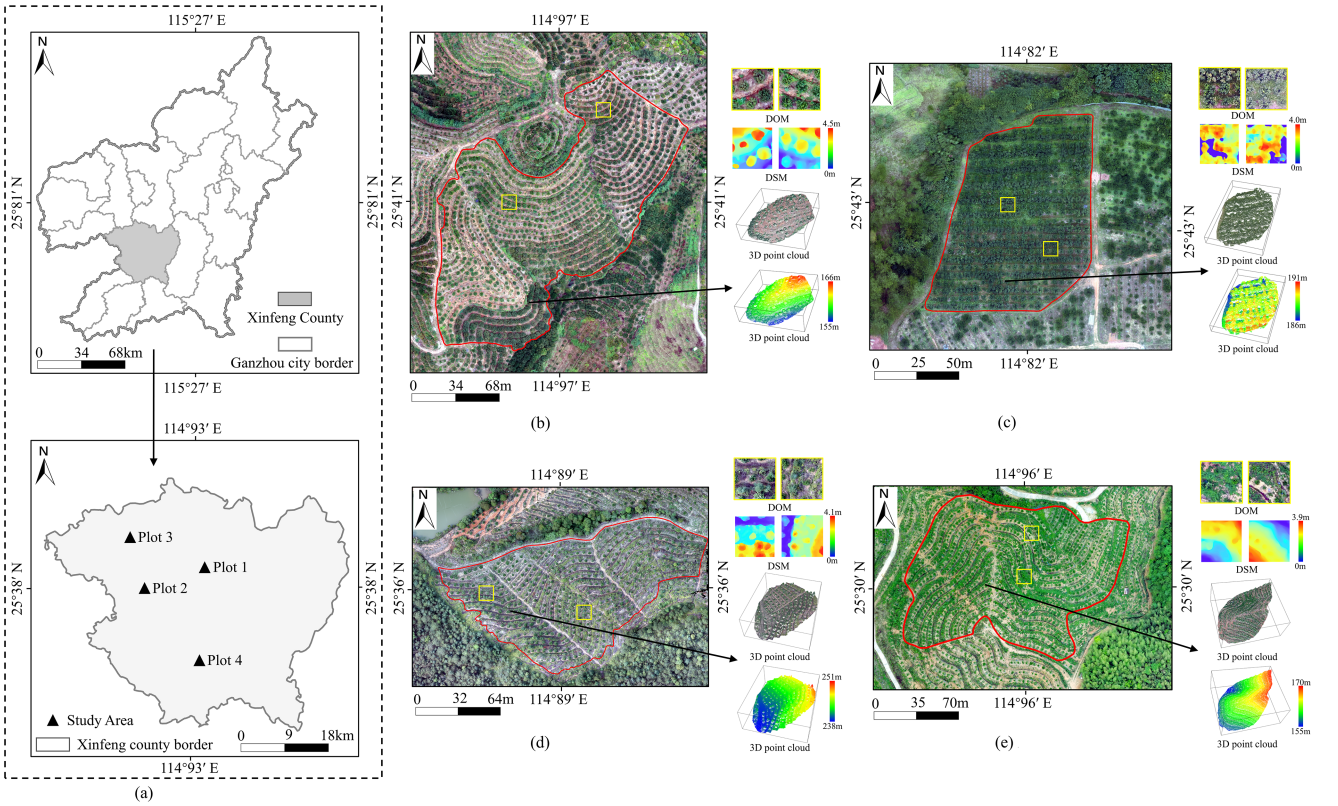

Fig. 5. Study area. (a) is the location of the study area. The subareas marked by red lines in (b), (c), (d), and (e) are Plot 1, Plot 2, Plot 3, and Plot 4, respectively. The subgraphs located to the right of (b), (c), (d), and (e) are DOM, DSMs, 3-D point clouds with RGB textures, and 3-D point clouds rendered by height, respectively.

We used photogrammetric software called Agisoft Photoscan [44] to perform aerial triangulation for generating high-resolution DOM and DSM with a spatial resolution of 4 cm/pix. In addition, pixel-level dense point clouds were generated to characterize the geometric morphology of the citrus tree canopy surface. Generally, a certain distance between citrus trees is observed in the planting area, and in most cases, each citrus tree is surrounded by an open space, which is lower in height than the citrus tree. In terms of the principle of the cloth simulation filter [45], this algorithm is particularly suitable to separate the ground point cloud from the dense point cloud. Then, digital terrain models (DTMs) of Plot 1, Plot 2, Plot 3, and Plot 4 were generated by the Kriging interpolation algorithm. Subsequently, the CHMs of Plot 1, Plot 2, Plot 3, and Plot 4 can be obtained by subtracting the corresponding DTM from the DSM. As shown in Fig. 6, from a visual perspective, the geometric morphology of citrus tree canopies can be well characterized by CHMs without being affected by terrain fluctuations.

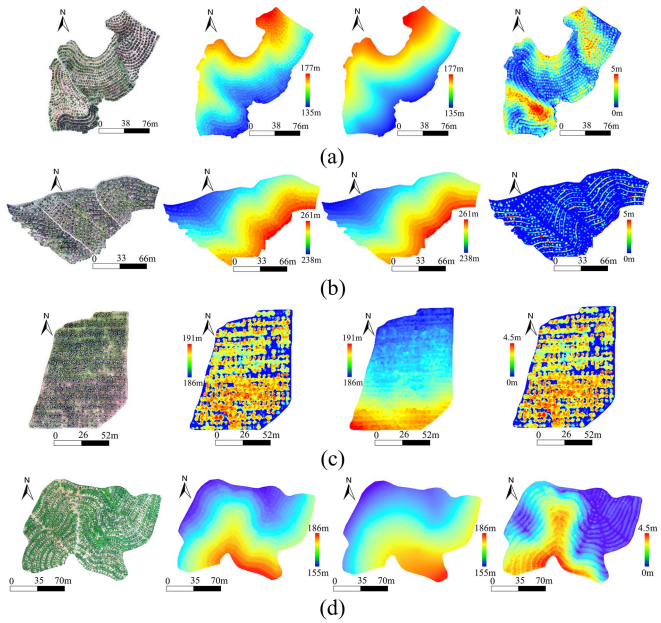

Fig. 6. DOM, DSM, DTM, and CHM of Plot 1, Plot 2, Plot 3, and Plot 4. The first to fourth columns represent the DOM, DSM, DTM, and CHM respectively.

C. Training and Validation

In the proposed network, given the use of several pretrained network modules through transfer learning, weight parameters can be shared without requiring a large amount of training sample data from scratch. In the training of the proposed network, we manually delineated samples from UAV photogrammetry-derived data using ArcGIS 10.8 software, requiring each sample to contain at least one citrus tree. Through data augmentation methods, such as rotation, flip, and affine transformation, the number of these samples has been increased by four times, generating a total of 25 000 samples with a patch size of 256×256.

TABLE I
TRAINING PLATFORM AND SETTINGS

Platform		Training Settings	
CPU	Intel (R) i5-12600kf@3.7GHz	Optimizer	AdamW
GPU	NVIDIA GeForce RTX 3060-12G	LR Policy	Poly
Memory	16GB	Loss Functions	BCE/Dice
DL Framework	Pytorch V1.12.0	Initial learning rate	0.0001
Compiler	PyCharm 2022.1.4	Betas	(0.9, 0.999)
Program	Python V3.7.15	Weight decay	0.001
Parallel computing	CUDA V11.3	Batch size	8
DL Accelerator	cuDNN V8.2.0	Epoch	60

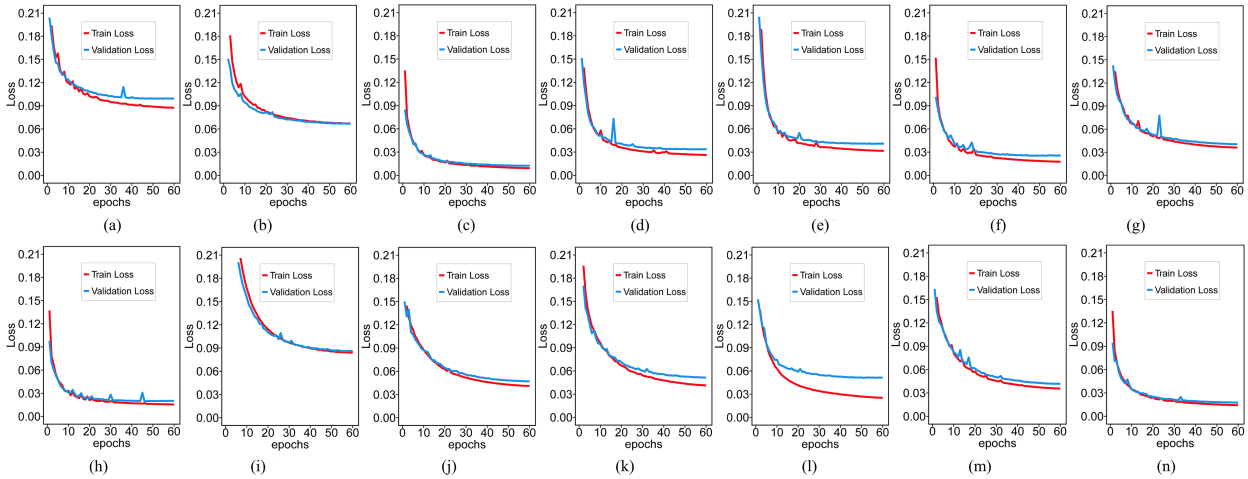

Fig. 7. Fine-tuning results of FCN, BiseNet-V2, U-Net, PSPNet, DANet, FPN, EfficientNet-V2-S, HRCNet_W48, DeepLab-V3, ResT-Tiny, Swin-Tiny, BeiT-Base, and CSwin-Tiny. (a) FCN. (b) BiseNet-V2. (c) U-Net. (d) PSPNet. (e) DANet. (f) FPN. (g) EfficientNet-V2-S. (h) HRCNet. (i) DeepLab-V3. (j) ResT-Tiny. (k) BeiT-Base. (l) Swin-Tiny. (m) CSwin-Tiny. (n) Proposed.

One-fifth of these samples are test datasets, and the remaining samples are training and validation sets.

The proposed network was trained in a deep learning framework named PyTorch based on the Python compiler. AdamW and cuDNN were used to optimize and accelerate model training, respectively. In addition, a learning rate decay strategy, mathematically expressed in (8), was conducted to accelerate the convergence of network training. The training platform and settings are detailed in Table I, in which Betas are the two momentum parameters in the AdamW algorithm

$$\text{learn\_rate} = \text{init\_learn\_rate} \times \left(1 - \frac{\text{iter}}{\text{max\_iter}}\right)^{\text{power}} \quad (13)$$

where learn_rate represents the learning rate, init_learn_rate represents the initial learning rate, iter represents the number of iterations, and max_iter represents the maximum number of iterations. In this study, the power is set to 0.9.

D. Comparative Analysis of Popular Modules

In this section, to prove the effectiveness of the modules in the popular deep networks (e.g., CNN, EfficientNet-V2, and CSwin Transformer) for citrus tree canopy segmentation, classic CNN models (including FCN [7], U-Net [8], FPN [46], PSPNet [34], DeepLab-V3 [9], BiseNet-V2 [47], DANet [11], and HRCNet [35]) and transformer models (including SwinT [19], CSwinT [20], Beit [18], and ResT [17]) were selected to evaluate the performance of citrus tree canopy segmentation. Among them, DANet, PSPNet, and DeepLab-V3 all use ResNet-50 [41] as the backbone. In addition, to restore the original image resolution, a decoder called FPNHead was introduced into the networks (e.g., EfficientNet-V2, FPN, SwinT, CSwinT, and ResT) for further evaluation of feature extraction. In addition, these networks performed a fine-tuning operation using the training samples to satisfy the application of citrus tree segmentation. The fine-tuning results are shown in Fig. 7. Whether trained or validated, the objective loss function of the proposed network can quickly converge and stabilize at a lower value. Therefore, the proposed network is more suitable for citrus tree canopy segmentation compared with the most popular deep networks, such as FCN, BiseNet-V2, and PSPNet.

The experimental results are given in Table II. By comparison, for this metric (i.e., mIoU), FCN and BiseNet-V2 using multiple convolutional layers have the lowest scores, with 86.41% and 86.95%, respectively. U-Net uses skip connections to fuse low-level high-resolution feature maps, with an accuracy 2.29% higher than that of FCN. The accuracy of FPN, PSPNet, and DeepLab-V3 using multiscale feature fusion was considerably improved in comparison with FCN, with a maximum accuracy improvement of 3.84%. Although the use of the attention mechanism in DANet and HRCNet_W48 can improve the accuracy of citrus tree canopy segmentation, its effectiveness remains

TABLE II
COMPARISON OF SEGMENTATION PERFORMANCE BETWEEN CNN AND TRANSFORMER ON THE CANOPY DATASET

Method	OA	F_1	mIoU	Para(M)	FLOPs(G)
FCN [7]	0.9508	0.8723	0.8641	15.90	160.97
BiseNet-V2 [47]	0.9607	0.8768	0.8695	3.62	25.75
U-Net [8]	0.9693	0.8982	0.8870	13.40	248.23
PSPNet [34]	0.9691	0.9078	0.8973	72.46	347.40
DANet [11]	0.9694	0.9084	0.8981	70.94	405.02
FPN [46]	0.9696	0.9096	0.8991	53.71	105.11
EfficientNet-V2-S [40]	0.9690	0.9102	0.9005	26.76	222.62
HRCNet_W48 [35]	0.9702	0.9111	0.9008	62.71	187.38
DeepLab-V3 [9]	**0.9712**	0.9125	0.9025	65.50	347.04
ResT-Tiny [17]	0.9677	0.9041	0.8935	18.50	211.02
Swin-Tiny [19]	0.9680	0.9047	0.8941	36.80	243.52
BeiT-Base [18]	0.9685	0.9066	0.8960	102.99	424.93
CSwin-Tiny [20]	0.9708	**0.9136**	**0.9033**	30.31	240.75

The bold entities indicate the maximum value.

TABLE III
COMPARISON OF DIFFERENT SIZES OF NETWORK BACKBONE STRUCTURES

CNN	Transformer	OA	F_1	mIoU
EfficientNet-V2-S	CSwin-Tiny	0.9735	0.9498	0.9346
EfficientNet-V2-M	CSwin-Tiny	0.9739	0.9506	0.9355
EfficientNet-V2-L	CSwin-Tiny	**0.9746**	**0.9519**	**0.9372**
EfficientNet-V2-L	CSwin-Small	0.9731	0.9508	0.9360
EfficientNet-V2-L	CSwin-Base	0.9730	0.9511	0.9364
EfficientNet-V2-L	CSwin-Large	0.9726	0.9494	0.9350

The bold entities indicate the maximum value.

TABLE IV
COMPARISON OF ABLATION STUDIES

Method	CAFM	CA	$Loss_{aux}$	*OA*	*F*1	*mIoU*
Proposed		✓	✓	0.9684	0.9069	0.8973
	✓		✓	0.9695	0.9138	0.9036
	✓	✓		0.9701	0.9194	0.9085
	✓	✓	✓	**0.9710**	**0.9237**	**0.9116**

The bold entities indicate the maximum value.

slightly lower than the DeepLab-V3 network. EfficientNet-V2-S using the FPNHead decoder also achieved a score of 90.05%, which is 0.20% lower than the highest score of DeepLab-V3 in the CNN-based models. Therefore, multiscale fusion and attention mechanisms can remarkably improve citrus tree canopy segmentation performance. In addition, transformer models with multiscale feature fusion and attention mechanisms, such as ResT-Tiny, Beit-Base, and Swin-Tiny, perform well in terms of mIoU but slightly lower than CNN-based models. This can be attributed to the transformer model's inherent limitation in capturing fine-grained local information of the canopy, which affects its ability to accurately delineate canopy boundaries. Among the popular transformer networks, CSwin-Tiny has the highest mIoU score, indicating that the LePE in the CSwin transformer module efficiently enhances the ability of CSwin-Tiny to extract local features.

Given the addition of global attention mechanisms, the above-analyzed networks, such as DANet with the multiscale feature fusion module, can perform better in terms of mIoU, but they also have higher computational complexity than networks without mechanism modules. Compared with CNN-based networks with full attention mechanisms, transformer modules based on local attention mechanisms (such as the cross-shaped attention mechanism used in CSwin transformer) can considerably reduce the number of network parameters and computational complexity. Although the computational complexity of CSwin transformer is not minimal compared with other networks, such as ResT-Tiny, its segmentation accuracy is considerably higher in terms of metrics, such as OA, *F*1, and mIoU, than other networks except for DeepLab-V3. In addition, Table II presents that the CNN network for local semantic information extraction, namely, EfficientNet-V2-S based on the FPNHead decoder, is the lightest weight network in addition to FCN, U-Net, and BiseNet-V2; nevertheless, its accuracy is much higher than FCN, U-Net, and BiseNet-V2. Therefore, the modules from networks, such as EfficientNet-V2 and CSwin, enable effective segmentation of citrus tree canopies.

In addition, we evaluated the performance of different sizes of network backbone structures, such as EfficientNet-V2 and CSwin. Three sizes of CNN-based modules, namely, EfficientNet-V2-S, EfficientNet-V2-M, and EfficientNet-V2-L, and four sizes of transformer-based modules, namely, CSwin-Tiny, CSwin-Small, CSwin-Base, and CSwin-Large, were designed to perform citrus tree canopy segmentation. The experimental results are shown in Table III. As the EfficientNet-V2 deepens, the number of module parameters increases sharply, and the performance improves slightly but not remarkably. Notably, deeper CSwin transformer networks demonstrate a reduction in tree crown segmentation performance. Therefore, the above comparative analysis suggests that the proposed network combining EfficientNet-V2-S as the CNN branch and CSwin-Tiny as the transformer branch is a tradeoff between network parameters and segmentation accuracy.

E. Ablation and Studies

To verify the effectiveness of CAFM, CA, and $Loss_{aux}$ on model performance, we conducted ablation studies with the same 2-D data. The comparisons of ablation studies are given in Table IV. As shown in the table, the proposed network with CAFM, CA, and $Loss_{aux}$ outperforms other configurations

TABLE V
COMPARISON OF SEGMENTATION PERFORMANCE OF 2-D RGB IMAGERY OR 3-D DATA USED

Data	Method	OA	F_1	mIoU
2D	FCN	0.9508	0.8723	0.8641
	U-Net	0.9693	0.8982	0.8870
	EfficientNet-V2-S	0.9690	0.9102	0.9005
	HRCNet_W48	0.9702	0.9111	0.9008
	DeepLab-V3	**0.9712**	0.9125	0.9025
	CSwin-Tiny	0.9708	0.9136	0.9033
	Proposed	0.9710	**0.9237**	**0.9116**
3D	FCN	0.9537	0.9046	0.8816
	U-Net	0.9591	0.9241	0.9106
	EfficientNet-V2-S	0.9632	0.9341	0.9159
	HRCNet_W48	0.9668	0.9409	0.9243
	DeepLab-V3	0.9685	0.9394	0.9225
	CSwin-Tiny	0.9711	0.9407	0.9231
	Proposed	**0.9735**	**0.9498**	**0.9346**

The bold entities indicate the maximum value.

TABLE VI
COMPARISON BETWEEN 2-D AND 3-D FOR PLOTS 1–4

Plot #	Data	OA	Precision	Recall	F_1	mIoU
Plot 1	2D	0.9639	0.9217	0.9394	0.9305	0.9112
	3D	0.9694	0.9361	0.9456	0.9408	0.9239
Plot 2	2D	0.9792	0.9101	0.9142	0.9122	0.9088
	3D	0.9827	0.9095	0.9305	0.9199	0.9162
Plot 3	2D	0.9674	0.9568	0.9605	0.9586	0.9341
	3D	0.9737	0.9660	0.9672	0.9666	0.9465
Plot 4	2D	0.9480	0.8249	0.8698	0.8467	0.8554
	3D	0.9643	0.9009	0.9303	0.9153	0.8996

across all metrics, suggesting that the feature fusion model, attention mechanism, and auxiliary loss function can help improve the performance of citrus tree canopy segmentation. Individually removing CAFM, CA, and Loss_{aux} resulted in decreases of 1.43%, 0.80%, and 0.31% in the mIoU accuracy of the model, suggesting the substantial contribution of CAFM in improving model performance, followed by CA and Loss_{aux}. The improvement can be explained as follows: On the one hand, CAFM can effectively reduce the loss of information in the process of CNN and transformer feature fusion while considering local and global semantic information for citrus tree canopy segmentation; on the other hand, the superposition of CA enhances the proposed model's perception of the canopy boundary, thereby improving the overall accuracy of citrus tree canopy segmentation. Meanwhile, in the model decoding stage, the CA module can further extract the position information of citrus tree canopies during the restoration of the original image resolution, thus helping alleviate the loss of spatial information. In addition, the use of Loss_{aux} is conducive to learning more effective semantic representation in the training stage.

F. Comparison of 2-D and 3-D Data for Citrus Tree Canopy Segmentation

To overcome the influence of terrain relief and complex backgrounds (e.g., low weeds), in this study, additional data, such as CHM, were selected to characterize the 3-D geometric structure of citrus tree canopies. To verify the effectiveness of performance improvement because of the CHM used, in this section, we investigate the effect of 3-D data on citrus tree canopy segmentation and conduct experiments in three plots with varying terrains and complex backgrounds. The experimental results are given in Table V.

As shown in Table V, whether 2-D or 3-D data are used, the proposed network can considerably perform better for citrus tree canopy segmentation than when only CNN or transformer is used. In terms of metrics, such as mIoU, the proposed network is 4.75% and 2.46% higher in 2-D data than FCN and U-Net, respectively. In addition, the proposed 3-D network is at least 0.25%, 2.61%, and 2.30% higher than the 2-D data used in terms of OA, $F1$, and mIoU, respectively. The comparison results indicate that adding CHM can effectively improve the performance of citrus tree canopy segmentation, as all the compared networks based on 3-D data showed better statistical results than those based on 2-D data.

To verify further the performance of citrus tree canopy segmentation using 3-D data under complex backgrounds, such as terrain relief, surrounding shrubs, or weeds, we provided a more detailed comparison in Plot 1, Plot 2, Plot 3, and Plot 4. The segmentation results of citrus tree crowns in various regions are shown in Fig. 8, and the experimental results are shown in Table VI. In terms of the mIoU of Plot 1, compared with 2-D data, using 3-D data reduced the impact of terrain fluctuations, resulting in an average increase of 1.27%. In contrast to Plot 1, although Plot 2 is also an area with considerable terrain fluctuations, remarkable interclass differences are observed in surface coverage and a considerable distance between citrus tree canopies, resulting in the addition of citrus tree canopy height information having minimal effect on the improvement of citrus tree canopy segmentation performance, with mIoU only increasing by 0.74%. For Plot 3, using 3-D data can remarkably improve the accuracy (1.24% higher mIoU) of citrus tree canopy segmentation in local areas containing low shrubs and weeds, which are similar in visible-light spectrum and texture to citrus tree canopies. Especially for Plot 4, due to the presence of large terrain undulations, dense shrubs and weeds, and adhesive tree crowns, it is very difficult to distinguish between the tree canopy and the background. The mIoU using 2-D data is only 85.54%, but the mIoU using 3-D data is 89.96%, which is 4.42% higher than that using 2-D data. The above quantitative analysis reveals that, in artificial forest areas with large terrain fluctuations, complex canopy backgrounds, and severe canopy adhesion, adding CHM to 2-D RGB imagery can effectively improve the performance of citrus tree canopy segmentation.

For an intuitive comparison, a visualization evaluation was also conducted to evaluate the performance of citrus tree canopy segmentation. Several representative segmented patches were selected in Fig. 9. As shown in Fig. 9(a) and (b), the boundaries between cohesive citrus tree canopies can be clearly delineated by using 3-D data. In addition, Fig. 9(c)–(g) shows that the proposed network based on 2-D RGB imagery cannot accurately distinguish shrubs or weeds, and misclassifications and omissions are observed. By contrast, the proposed network with CHM can help accurately delineate the boundaries of citrus tree canopies and is less likely to miss segmented patches.

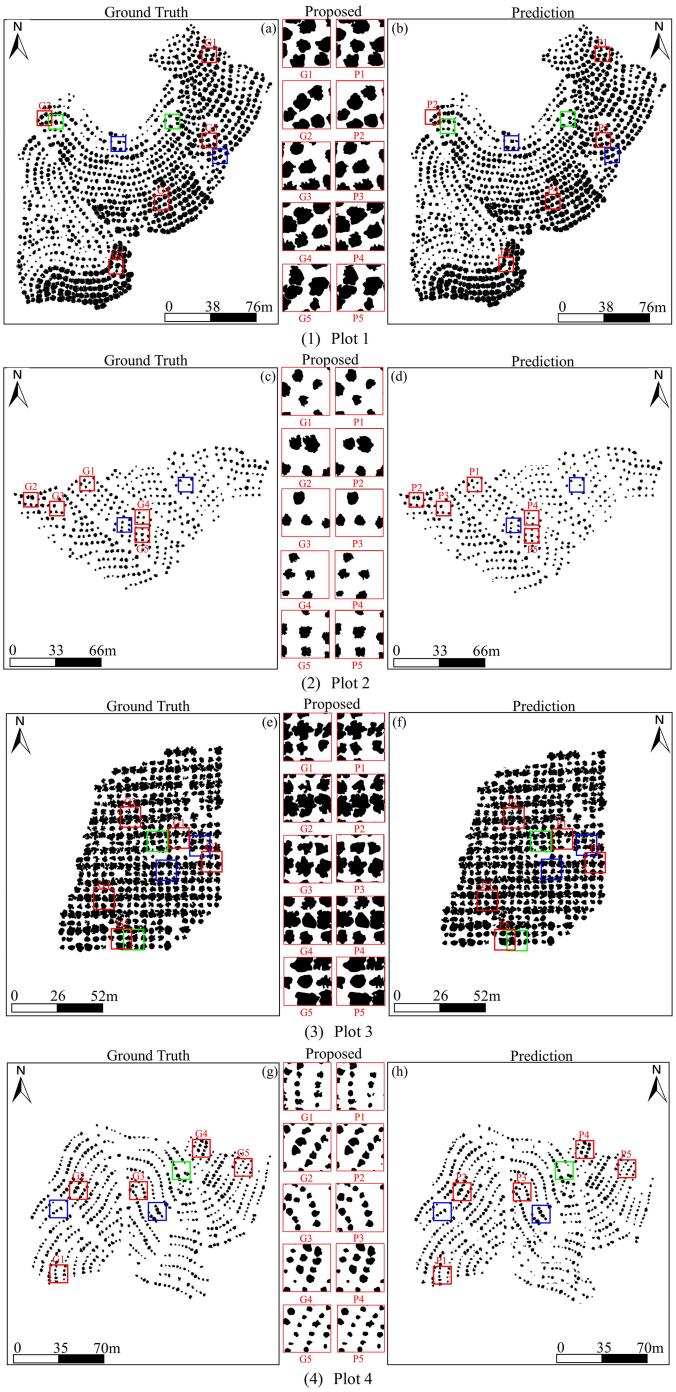

Fig. 8. Segmentation results obtained by the proposed method. (1) (2), (3), and (4) are the ground truth and prediction for Plot 1, Plot 2, Plot 3, and Plot 4, respectively. (a), (c), (e), and (g) are the ground truth; (b), (d), (f), and (h) are the predictions of the proposed method. The blue box represents the position of the comparison image for Experiment F, and the green box represents the position of the comparison image for Experiment G.

G. Comparison With Other State-of-the-Art Networks

To verify the overall effectiveness of the proposed parallel fusion neural network further, state-of-the-art networks, such as SETR_PUP [16], TransUNet [36], TransFuse [38], and CCTNet [48], were selected to perform citrus tree canopy segmentation, and the statistical experimental results are given in Table VII.

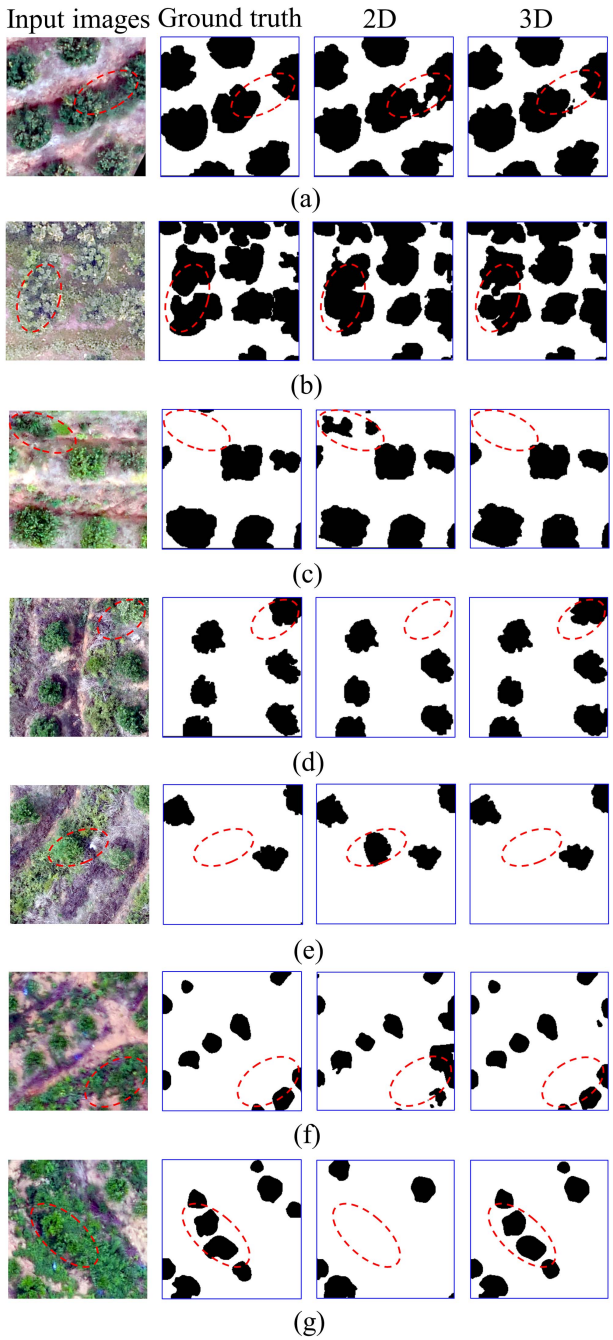

Fig. 9. Visualization evaluation using 2-D and 3-D data. (a)-(g) are seven representative examples selected in Plots 1-4 to display the visualization effect. Note that the significantly contrasting areas are marked with red circles.

The proposed network with the highest mIoU score performs better than the four state-of-the-art networks fusing CNN and transformer modules. SETR_PUP uses an attention mechanism in the encoder to extract image features and employs convolution for decoding. However, compared with CNNs, it has limitations in capturing local feature information. TransUNet uses a serial structure to fuse CNN and transformer, which cannot effectively retain local and global contextual information extracted by CNN and transformer modules, respectively, and, therefore, cannot accurately detect the edges of citrus tree canopies. Although

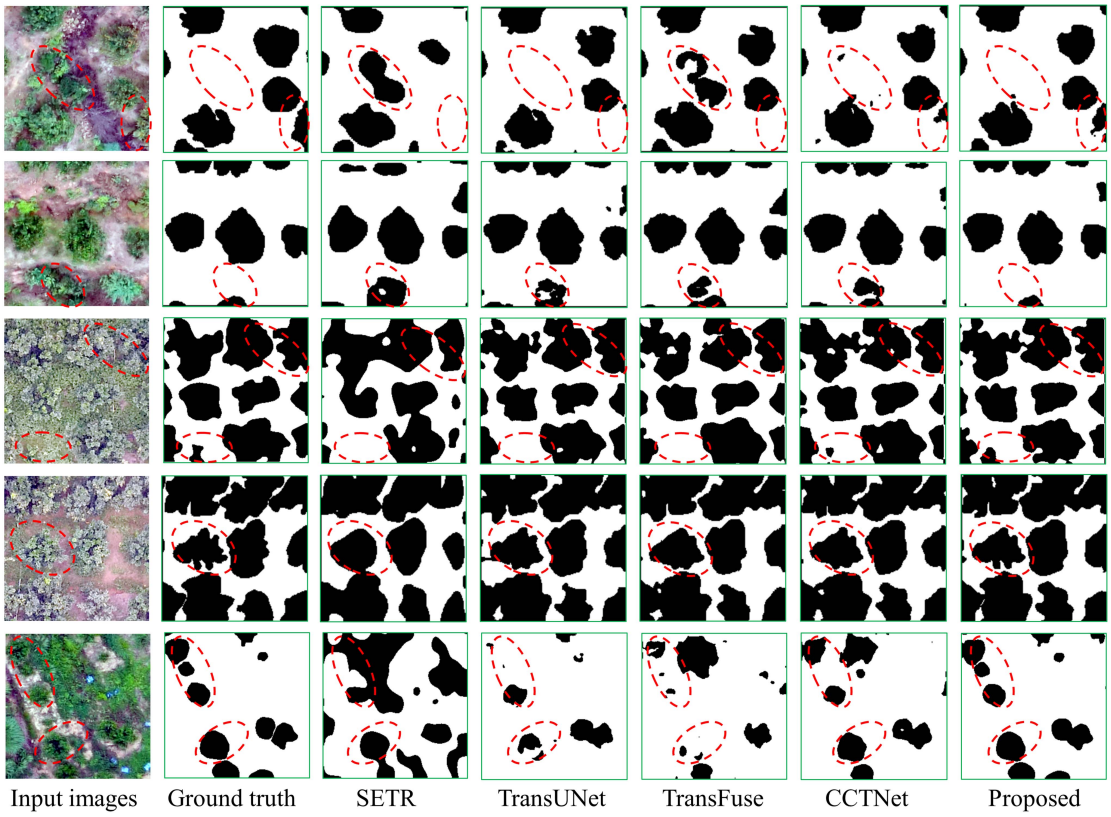

Fig. 10. Visualization comparison of the state-of-the-art networks and the proposed network.

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS

Method	OA	F_1	mIoU	Para(M)	FLOPs(G)
SETR_PUP [16]	0.9386	0.8842	0.8561	86.07	192.39
TransUNet [36]	0.9701	0.9404	0.9236	37.39	112.22
TransFuse [38]	0.9690	0.9384	0.9211	70.34	191.46
CCTNet [48]	0.9706	0.9309	0.9171	62.95	178.92
Proposed	**0.9735**	**0.9498**	**0.9346**	50.25	193.16

The bold entities indicate the maximum value.

TransUNet can alleviate the problem of insufficient spatial information during decoding to some extent by using skip connections, this serial architecture could damage the respective characteristics of the CNN and transformer. In contrast to the structure of TransUNet, TransFuse and CCTNet apply a parallel structure to fuse CNN and transformer modules and design corresponding fusion modules for features in each branch. Although the feature fusion module can combine the advantages of CNN and transformer modules, it cannot effectively retain the position information of each canopy, resulting in inaccurate canopy boundary delineation. Compared with the four state-of-the-art networks, the proposed network not only inherits the advantages of CNN and transformer modules but also extracts the position information, thereby accurately identifying citrus tree canopies.

In addition, several patches were selected to exhibit the visual segmentation results in Fig. 10. The proposed network can perform much better than the state-of-the-art networks because citrus tree canopies can be accurately identified, and boundaries can be accurately delineated.

V. CONCLUSION

In this study, we proposed a parallel fusion neural network considering local and global semantic information for citrus tree canopy segmentation from UAV photogrammetry-derived 3-D data. The proposed network, coupling CNN and transformer (i.e., EfficientNet-V2 and CSwin), can address the problem in which traditional semantic segmentation methods cannot effectively retain local boundary details and global contextual information of citrus tree canopies. Specifically, a module named CAFM was explored to fuse features obtained by CNN and transformer modules, and a CA mechanism was utilized to extract the position information of citrus tree canopies and further optimize their boundaries. In addition, to solve the problem of insufficient

2-D data to characterize the geometric structure of citrus tree canopies in complex terrain and backgrounds, the CHM of citrus tree canopies was added to the proposed network to improve the feature extraction performance by adding input dimensions. Transfer learning based on EfficientNet-V2 and CSwin also provided the opportunity to improve the performance of citrus tree canopy segmentation remarkably without the requirement of training from scratch. Compared with the state-of-the-art networks, the proposed method considerably performs better than networks based only on CNN or transformer models, and shows the best citrus tree canopy segmentation results (e.g., the highest mIoU score of 93.46%) in terms of several metrics.

Although we have made remarkable improvements by using 3-D data combining 2-D true-color RGB imagery and CHM in this study, further performance improvements are still needed to eliminate the effect of shrubs that are highly similar to citrus trees. In future research, we will attempt to collect multispectral data from UAVs to filter noncitrus trees by increasing the gap between citrus and noncitrus trees based on vegetation spectral differences.

REFERENCES

[1] O. Satir, S. Berberoglu, E. Akca, and O. Yeler, "Mapping the dominant forest tree distribution using a combined image classification approach in a complex eastern Mediterranean basin," *J. Spatial Sci.*, vol. 62, no. 1, pp. 157–171, Jul. 2016.

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

[3] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," *Remote Sens.*, vol. 15, no. 7, Mar. 2023, Art. no. 1860.

[4] H. He, C. Li, R. Yang, H. Zeng, L. Li, and Y. Zhu, "Multisource data fusion and adversarial nets for landslide extraction from UAV-photogrammetry-derived data," *Remote Sens.*, vol. 14, no. 13, Jun. 2022, Art. no. 3059.

[5] B. Niu, Q. Feng, B. Chen, C. Ou, Y. Liu, and J. Yang, "HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery," *Comput. Electron. Agr.*, vol. 201, 2022, Art. no. 107297.

[6] H. Yurtseven, M. Akgul, S. Coban, and S. Gulci, "Determination and accuracy analysis of individual tree crown parameters using UAV based imagery and OBIA techniques," *Measurement*, vol. 145, pp. 651–664, 2019.

[7] 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.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[11] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.

[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[15] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021, pp. 1–22.

[16] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[17] Q. Zhang and Y.-B. Yang, "Rest: An efficient transformer for visual recognition," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15475–15485, 2021.

[18] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," 2021, *arXiv:2106.08254*.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[20] 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.

[21] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retrieval*, vol. 7, pp. 87–93, 2018.

[22] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[23] S. Bose, R. S. Chowdhury, D. Pal, S. Bose, B. Banerjee, and S. Chaudhuri, "MultiScale probability map guided index pooling with attention-based learning for road and building segmentation," 2023, *arXiv:2302.09411*.

[24] C. Qiu et al., "Transferring transformer-based models for cross-area building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4104–4116, 2022.

[25] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3585.

[26] 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, Nov. 2018, Art. no. 736.

[27] E. Guirado et al., "Mask R-CNN and OBIA fusion improves the segmentation of scattered vegetation in very high-resolution optical sensors," *Sensors-Basel*, vol. 21, no. 1, Jan. 2021, Art. no. 320.

[28] 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.

[29] 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, Jul. 2021, Art. no. 2721.

[30] 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.

[31] 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.

[32] J. A. C. Martins et al., "Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3054.

[33] 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, Jan. 2021.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[35] 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.

[36] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[37] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[38] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 14–24.

[39] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[40] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[42] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.
[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[44] Agisoft LLC, Agisoft PhotoScan User Manual, Jul. 25, 2019. [Online]. Available: https://www.agisoft.com/pdf/photoscan-pro_1_4_en.pdf
[45] W. Zhang et al., "An easy-to-use airborne LiDAR data filtering method based on cloth simulation," *Remote Sens.*, vol. 8, no. 6, Jun. 2016, Art. no. 501.
[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[47] 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.
[48] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 9, Apr. 2022, Art. no. 1956.

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

He is currently a Full Professor with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China, and also with the Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang, China. His research interests include photogrammetry and remote sensing, image processing, and machine learning.

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

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

Yuanping Xia received the Ph.D. degree in cartography and geographical information engineering from the China University of Mining and Technology, Xuzhou, China, in 2020.

He is currently a Full Professor with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China, and also with the Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang, China. His research interests include remote sensing, land environment, and disaster monitoring.

Min Chen received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2014.

He is currently an Associate Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. His research interests include photogrammetry and remote sensing.

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.