Scale Pyramid Graph Network for Hyperspectral Individual Tree Segmentation

Yaqian Long, Songxin Ye, Liqiong Wang, Weixi Wang, Xiaomei Liao, and Sen Jia, *Senior Member, IEEE*

***Abstract*— Unmanned aerial vehicle (UAV) hyperspectral imaging offers an efficient and cost-effective way to map tree species at the individual tree levels. Conventional methods mostly rely on large samples of natural RGB images of tree crowns, lacking the ability to distinguish species, particularly for trees with overlapping crowns. This study proposed a novel scale pyramid graph network (SPGN) for instance segmentation that can simultaneously apply pixel-level (node) classification for discriminating species and edge prediction for delineating individual trees. Based on a graph-in-graph (GiG) convolution, we built a scale pyramid module (SPM) that extracts multiscale features at pixels, superpixels, and subgraph levels to aggregate the over-segmented superpixels into the same species and the same tree. We also proposed an innovative concept of subgraph positional encoding (SPE) to represent the natural spatial relationship of graph-structured data. The SPGN method was evaluated in a case study involving eleven subtropical broadleaf species under an urban environment in south China. The accuracy of species classification achieved 93%, and the area under the curve (AUC) of individual tree segmentation reached 0.96. Compared with state-of-the-art methods such as DeepForest, Detectree2, and segment anything model (SAM), SPGN presented fewer errors in tree detection and outperformed in instances of crown overlaps. Ablation studies proved the effectiveness of SPM and SPE modules, which improved segmentation by 10% and classification by 7% in accuracy, respectively. The findings confirm the benefits of incorporating spatial context, such as crown textures and tree positional relationships, for species differentiation; in return, accurate species identification combined with spectral information assists the individual tree segmentation. This effective strategy can be potentially extended to a broader range of regions and forest types.**

Manuscript received 21 February 2024; revised 3 July 2024; accepted 19 July 2024. Date of publication 5 August 2024; date of current version 15 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903702; in part by the National Natural Science Foundation of China under Grant 62271327 and Grant 62001303; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290 and Grant 2023A1515012053; in part by Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005; and in part by the Research Team Cultivation Program of Shenzhen University under Grant 2023JCT002. *(Corresponding author: Sen Jia.)*

Yaqian Long, Songxin Ye, Liqiong Wang, and Sen Jia are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: longyaqian@szu.edu.cn; yesongxin2021@email.szu.edu.cn; wangliqiong2023@email.szu.edu.cn; senjia@szu.edu.cn).

Weixi Wang is with the College of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China (e-mail: wangwx@szu.edu.cn).

Xiaomei Liao is with the College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518060, China (e-mail: liaoxm@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3439094

***Index Terms*— Hyperspectral image (HSI), individual tree segmentation, scale pyramid graph network (SPGN).**

I. Introduction

THE tree species at individual tree level is fundamental information in forest inventory [1]. It characterizes the structure, biodiversity, and turnover of forests and facilitates monitoring the health conditions of forest ecosystems [2]. Furthermore, since tree species have varying carbon sequestration capabilities, identifying individual tree species allows for an accurate estimation of carbon storage and thus enables the possibility of understanding the differential responses of trees to climate change [3]. Compared to field surveys, remote sensing offers an efficient and potentially cost-effective means for mapping tree species [4], [5], [6]. Over the past few decades, however, remote sensing-based methods have predominantly focused on community-scale, pixel-level classification, with only a few studies being tree-centric [7]. Yet in the few tree-centric studies, the majority has been the segmentation of individual tree crowns using meter-level or submeter-level optical images [8], [9]. For instance, satellite images such as WorldView and QuickBird [10] have been employed to locate the scattered trees in African savanna drylands for the estimation of carbon stocks [11]. Based on the National Ecological Observation Network (NEON) project [12], ecologists have also established a benchmark database encompassing millions of hand-annotated tree crowns collected from airborne flight campaigns and a tree detection tool named "DeepForest" [13]. This tool has been applied across multiple forest types and geographical areas, such as temperate deciduous forests, boreal forests, and Mediterranean olive groves [14], [15], [16], yielded accurate identification of individual trees, and revealed the great potential and advantages of high-resolution images. However, these high-resolution images typically contain limited spectral information, constraining their use for distinguishing tree types, especially at species level [10].

Compared to RGB and multispectral data, hyperspectral imaging contains much more spectral information. Different species exhibit unique spectral signatures due to variations in chemical compositions and physical structures of their leaves and canopies [17]. Hyperspectral imaging excels at capturing these spectral signatures, with narrow and contiguous spectral bands enhancing critical identifiers, such as the chlorophyll peak (~550 nm) and the near-infrared (NIR) plateau (~690 nm) [18]. With wavelength-dependent analysis,

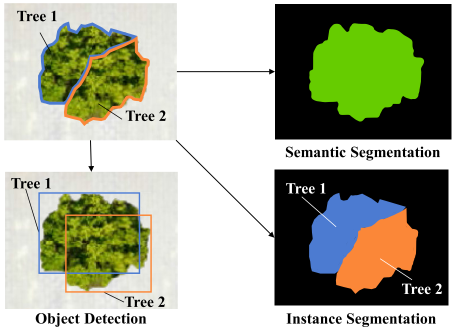


Fig. 1. Three different segmentation techniques: object detection, semantic, and instance segmentation for tree identification.

hyperspectral imaging is further able to extract subtle spectral differences among certain species and to enhance the connection between reflectance bands and optical properties of leaf tissue or canopy structure [19]. During the past two decades, airborne imaging spectroscopy has demonstrated robust performance in community-scale tree species identification, as evidenced by high classification accuracy in a variety of forest types [3], [20], [21]. Unmanned aerial vehicles (UAVs), also known as drones, can capture hyperspectral images at millimeter spatial resolution. As an alternative to traditional aircraft-based platforms, the UAV has made tree mapping more affordable and accessible, especially for small-scale projects or resource-constrained areas. The spatial resolution of images acquired by drones can surpass that of aircraft airborne imaging [22]. With such high spectral and spatial resolution, UAV-HSI alone should be sufficient for individual tree species classification [23]. However, labeling crown edges and species information in UAV images requires additional cost, labor, and expert knowledge; the scarcity of labeled hyperspectral data for training poses challenges in accurate tree segmentation.

Tree segmentation exclusively from UAV-hyperspectral images faces unique challenges. First, tree canopies can be vastly different in structure and appearance, making them difficult to distinguish in complex background environments [23]. Second, in densely vegetated areas, the crowns of trees often overlap each other and the gap between adjacent trees can be affected by shadow effects, both of which complicate the task of distinguishing the boundaries of individual tree crowns. Third, species such as broadleaf plants, can exhibit great spectral similarity, adding difficulty for their discrimination [24]. This is where channel attention mechanisms become important for the hyperspectral images with hundreds of spectral bands [25]. Channel attention focuses on relevant spectral bands that are most informative for distinguishing between similar species. By weighting the importance of each channel or spectral band associated with leaf compounds and canopy scale, these mechanisms can enhance the model's ability to focus on subtle spectral differences that are critical for accurate classification and segmentation in complex vegetative environments. In addition to spectral features, spatial features such as distinctive texture information can also be deterministic for individual tree segmentation and species classification.

Deep neural networks can effectively extract deep features and are well suited for handling complex tasks such as individual tree species identification [26]. Instance segmentation, as an advanced technique of image analysis in computer vision, aims to identify different objects ("instances") in images and delineate precise boundaries for the pixels of each object [27], [28], therefore providing a suitable tool for individual tree species mapping. While semantic segmentation is often applied to pixel-level classification of tree species, it cannot discriminate each tree as a unique instance. Instance segmentation combined the tasks of object detection and semantic segmentation (see Fig. 1). More importantly, it delineates the exact boundaries of the crown, which allows for measuring the crown width and area. Since object detection only generates bounding boxes of the individuals [13], the adjacent crowns would be easily confused with each other, hindering the accurate measure of the crown area which is a key parameter of tree properties. Since being proposed in 2017, the mask R-CNN has become a baseline model for instance segmentation. Although researchers have directly it applied to segment trees from RGB [15] and multispectral images [29], there are still several constraints. The top-view remote sensing images differ from conventional natural images which are mostly front view [30]. Remote sensing images cover a large spatial range and often exhibit characteristics such as being blurry and having low contrast. In addition, the ground targets in these images typically have unclear boundaries, making them difficult to identify through visual interpretation alone. Several networks have been proposed to address the unique characteristics of remote sensing images, including context aggregation network (CATNet) [30], semantic attention and scale complementary network (SEA-SCMB) [31]. These networks have primarily relied on RGB remote sensing images, constraining the exploration of rich spectral information available in hyperspectral images, which can significantly enhance the model's ability to classify and segment various materials and conditions on the Earth's surface. Recently, Fang et al. [32] proposed a module of spectral–spatial feature pyramid network (SS-FPN) to deal with the complex environment of remote sensing scenes and abundant spectral information. However, given that the approach was tested on man-made building targets, its applicability and performance in segmenting individual trees need to be further investigated.

This study proposed a novel network of instance segmentation for individual tree segmentation and species classification. The main contribution of this study is summarized below from three perspectives.

1) A scale pyramid structure, combined with a graph convolution network with reduced parameters, was built to extract representative features at three different levels of pixel, superpixel, and subgraph. This structure improved the classification accuracy of spectrally similar species and benefited tree segmentation, based on the ablation studies of this proposed module in performance assessments of the scale pyramid graph network (SPGN).
2) A novel method named subgraph positional encoding (SPE), which describes tree crowns' spatial relation based on graph structure and positional representation based on encoding and decoding, was proposed. The SPE showed improved separation of overlapping crowns

in densely vegetated areas, in terms of qualitative and quantitative assessments in individual tree segmentation tasks.

3) A series of experimental studies have demonstrated the high accuracy of UAV-hyperspectral images for the species classification of 11 broadleaf plants at the individual scale. The SPGN network delivered an increased mAP for tree segmentation and reduced efficiency, in comparison with methods such as mask R-CNN and SS-FPN.

The rest of the article is organized as follows. Section II summarizes the related work in image segmentation and graph convolution networks. Section III describes the proposed method of SPGN. The experimental results are introduced in Section IV. Finally, Section V concludes this article and provides some ideas for future research.

II. Related Work

This section clarifies the concepts of different segmentation techniques, each to be followed by a detailed description of representative methods.

A. Image Segmentation

Image segmentation is a technique for dividing the pixels in an image into multiple segments that represent different objects. Traditional pixel-based classification, while detailed, often lacks computational efficiency and can be prone to noise [33]. To address these issues, superpixel segmentation is often utilized. As a clustering method, it involves dividing an image into small clusters of adjacent pixels that share similar characteristics, such as color, brightness, and texture. These clustered pixels are also termed "superpixels." For example, normalized cut (NCut), introduced by Shi and Malik [34], measures similarity between and within groups to segment static images and motion sequences. Simple linear iterative clustering (SLIC), proposed by Achanta et al. [35], leverages the traditional k-means clustering method to efficiently generate superpixels. The SLIC algorithm utilizes a distance metric that combines color similarity and spatial proximity to form superpixels, which guarantees connectivity within the segmentation while requiring minimal memory demands.

Instance segmentation, which originated from computer vision, is an approach that aims to identify every individual object instance at the pixel level in an image [27]. It not only recognizes objects but also distinguishes among different instances within the same category, thereby providing precise boundaries for each instance. Two types of instance segmentation exist: two-stage and one-stage methods. The main difference between these types lies in the number of steps involved in object detection and pixel segmentation tasks. Two-stages methods have a two-step process: the first stage identifies object locations with bounding boxes, followed by a second stage that refines pixel-level classification within the identified regions. Pioneering studies utilized region-based convolutional networks (R-CNNs) to generate candidate object bounding boxes, namely, region proposals [36], and performed classification on each candidate. Subsequent work has focused on improving speed and accuracy. Variants such as fast R-CNN and mask R-CNN have proven to be simpler and more effective than many other networks. Mask R-CNN, in particular, has set a benchmark for instance-level recognition. While two-stage methods were once dominant in the field of instance segmentation, the step-by-step process fails to fully correlate the mutual information between object detection and instance segmentation. Furthermore, two-stage methods typically lack the high efficiency needed for real-time applications. This has led to the development of one-stage methods that combine both tasks into a single unified step. Proposals like the you only look once (YOLO) network [37] and the single shot multibox detector (SSD) [38] enable bounding box prediction and classification to be conducted concurrently during a single evaluation. More recently, transformer models [39] originally from natural language processing (NLP) [40], such as detection transformer (DETR) [41], have been introduced to reduce the computational burden for object detection.

For remote sensing images, instance segmentation is useful for capturing different ground targets, including vegetation and buildings. State-of-the-art networks from computer vision have been adopted and customized to accommodate the unique characteristics of remote sensing data and objects [32]. In addition, conventional instance segmentation methods face challenges when applied to trees with complex crown structures. Trees of different species can exhibit great similarity in spectral and spatial dimensions [3]. Moreover, the labeling for trees is scarce, restricting the effectiveness of deep neural networks in this domain. There remains a strong need to improve the quality of the tree segmentation network.

B. Graph Convolution Network

Graph convolutional networks (GCNs) constitute a powerful neural network architecture for processing data structured in the form of graphs [42]. This mechanism allows for capturing relationships and contextual information between nodes. Unlike traditional convolutional networks that are primarily designed for processing regular grid-like data (e.g., images), graph convolutions can effectively handle various types of irregular graph-structured data, such as social networks, knowledge graphs, and recommendation systems. Conventional GCNs [43] learn node representations by performing convolutions on the graph. The basic principle can be represented by the following equation:

$$H^{l+1} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{l}\Theta^{l} \tag{1}$$

where H^l and H^{l+1} represent the node representation matrix for the input and output of the lth layer, $\widetilde{A}$ is the adjacency matrix A with self-loops added, $\widetilde{D}$ is the diagonal degree matrix, and Θ^l is the weight matrix for the lth layer.

To reduce the complexity of the model, researchers simplify the architecture of GCNs by removing the nonlinear activation function and normalization steps. The computational efficiency of simplifying graph convolutional (SGC) networks was therefore improved [44]. Several GCNs have been proposed for hyperspectral images. For example, the multiscale dynamic GCN (MDGCN) establishes multiple input graphs

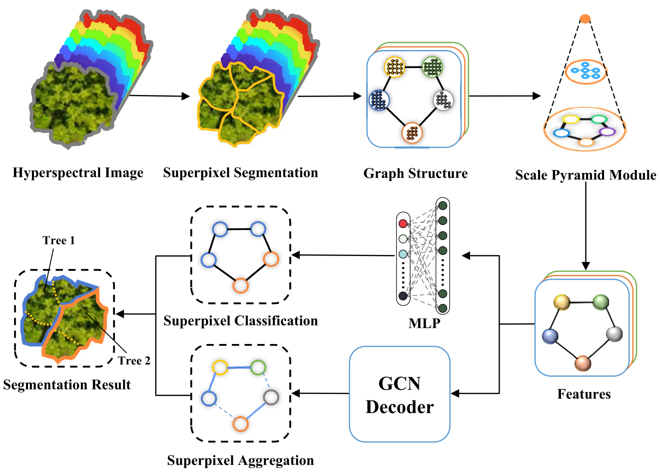


Fig. 2. Pipeline of the proposed method: hyperspectral data are imported to the preprocessing of superpixel segmentation, and segmented superpixel blocks are used for graph construction. The graph is then inputted into the SPM to extract features. The obtained features are separately fed into the MLP and GCN Decoder for superpixel classification and superpixel aggregation. The output results are fused to obtain the final segmentation result.

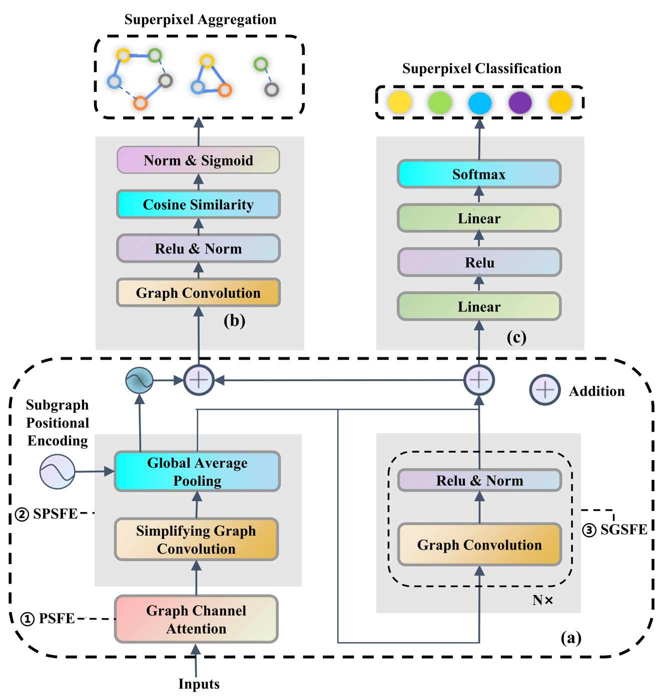


Fig. 3. Architecture of SPGN, which comprises three key modules. (a) SPM for feature extraction with three different scales, (b) GCN decoder for obtaining superpixel aggregation, and (c) MLP classifier for superpixel classification.

with different neighborhood scales to effectively leverage diverse spectral–spatial correlations at multiple scales, thereby achieving improved performance in hyperspectral image (HSI) classification [45]. The graph-in-graph convolutional network (GiGCN) [46] proposes a graph-in-graph (GiG) model and the corresponding GiGCN from the perspective of superpixels for HSI classification, demonstrating the effectiveness and feasibility of HSI classification with limited labeled samples.

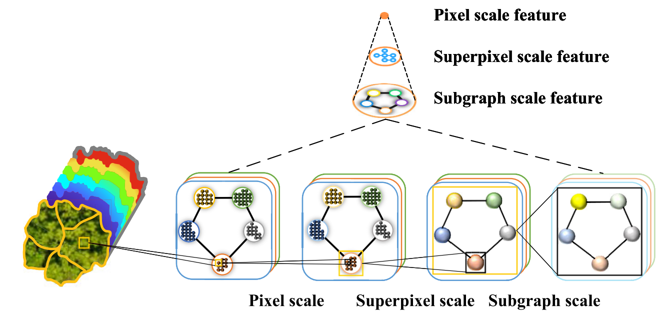


Fig. 4. Diagram of the scale pyramid: 1) Pixel features are extracted based on spectral difference at pixel level; 2) superpixel features are learned by clustering similar pixels; and 3) subgraph features are extracted by considering interaction information between adjacent subgraphs.

III. Methods

This study proposed a novel SPGN for individual tree species mapping. The pipeline of the proposed method is illustrated in Fig. 2. First, hyperspectral data undergo superpixel segmentation, followed by graph construction to obtain input for the network model. After the data input process, features are extracted through the scale pyramid module (SPM). The extracted features are then separately fed into the multilayer perceptron (MLP) [47] and GCN decoder to yield results for superpixel classification and superpixel aggregation. Finally, the outcomes of superpixel segmentation and superpixel aggregation are fused to obtain the ultimate segmentation result. The architecture of the network is depicted in Fig. 3. The network comprises three key modules: 1) SPM for feature extraction with three different scales; 2) GCN decoder for obtaining superpixel aggregation; and 3) MLP classifier for superpixel classification. The specific composition of each block will be elaborated in Sections III-A–III-C.

A. Graph Construction and Scale Pyramids for Improved Feature Extraction

The SPGN network has a graph structure that is inspired by a prior work named the GiGCN [46]. Before processing, hyperspectral data undergoes superpixel segmentation using the SLIC method, dividing the image into a series of superpixel blocks. Then, an internal graph is constructed to capture the inherent connections among the pixels within each superpixel block; simultaneously, an external graph is built to capture the neighboring relationships between different superpixel blocks. This multilevel graph structure lays a solid foundation for the subsequent in-depth extraction of spectral information.

An SPM was proposed for exploiting information from different levels constructed by the graph structure in the GiGCN network. As shown in Fig. 4, the SPM extracts features from three levels of scale: pixels, superpixels, and subgraphs. However, the graph structure only contains adjacency information, and it cannot capture the spatial context of positional information. To address this issue, we propose a module of SPE to further enhance the spatial information within the graph.

1) Pixel Scale Feature Extraction (PSFE): The first level of the scale pyramid, PSFE, was applied to perform channel

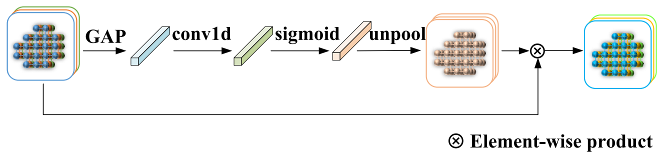


Fig. 5. Diagram of PSFE module. First, a 1-D vector is obtained through GAP on a 2-D image. Second, the 1-D convolution and sigmoid activation functions are applied, followed by unpooling to obtain an attention matrix with the same dimensions as the input. Finally, the attention matrix is element-wise multiplied with the input to obtain the output.

attention for graph convolution (CAGC) which was inspired by ECA-net [25]. Since tree species present unique spectral features across varying wavelength bands, PSFE facilitates band-specific weighting for species classification. The procedures of PSFE, illustrated in Fig. 5, include a global average pooling (GAP) that transforms a 2-D image into a 1-D vector, a 1-D convolution with a predefined kernel size (k) to maintain consistent learning parameters for all channels, a sigmoid activation, and an unpooling step that generates an attention matrix with the same dimensions as the input. The attention matrix is then element-wise multiplied with the input. PSFE results in reduced parameters in the convolutional network, thus increasing the network's efficiency.

2) SuperPSFE (SPSFE): The second level of the scale pyramid, SPSFE, was designed to learn feature representation at the superpixel scale. The SPSFE module comprises a simplifying graph convolution (SGC) layer and a GAP layer. In the process of feature extraction at the superpixel scale, the superpixel graph obtained from the previous layer, which contains new features, is first passed through multiple layers of graph convolutions to acquire the representation of each node in the superpixel graph. Subsequently, the information of all nodes in the superpixel graph is aggregated through GAP, resulting in the superpixel-scale features corresponding to the superpixel graph. After SPSFE, low-frequency information, such as texture features in homogeneous regions, is retained, while high-frequency information, such as noise, is greatly reduced. After the feature extraction process of each superpixel image via SPSFE, the corresponding features of the superpixel image can be obtained, and the features of all superpixel images are ultimately represented as the feature set $\mathbf{X}_g = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_S]^T$. This process can be represented as follows:

$$f^s_{*G}(\mathbf{X}_s, \mathbf{A}_s) = \mathrm{g}\left(\widetilde{\mathbf{X}}_s\right)$$
$$\widetilde{\mathbf{X}}_s = \left(\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\widetilde{\mathbf{A}}_s\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\right)^K \mathbf{X}_s\mathbf{W} \tag{2}$$

where $f^s_{*G}(\cdot)$ represents the function of SPSFE and $\mathrm{g}(\cdot)$ denotes GAP. Specifically, $\mathbf{W}$ is a trainable matrix in the SPSFE, $\widetilde{\mathbf{A}}_s$ is the adjacency matrix $\mathbf{A}_s$ with self-loops added, $\widetilde{\mathbf{D}}_s$ is the diagonal degree matrix, and K denotes neighbor hop.

3) Subgraph Scale Feature Extraction (SGSFE): The third level of the scale pyramid, SGSFE, aims to obtain neighborhood information between adjacent superpixels, by incorporating SPE into the GCN model. The neighborhood information is crucial for both node classification and edge prediction tasks. The feature extraction is based on a GCN, and the design of the network is shown in Fig. 6. An n layer of GCN

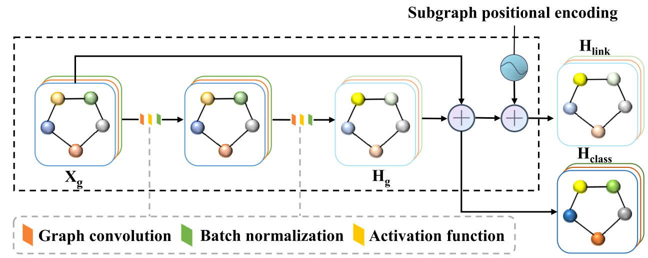


Fig. 6. Diagram of SGSFE module. Two layers of graph convolution are first developed to extract heterogeneous features, followed by BN and activation functions. Then, the features along with initial input are added to derive features for superpixel classification. Meanwhile, the positional information extracted using SPE is added to obtain features for superpixel aggregation.

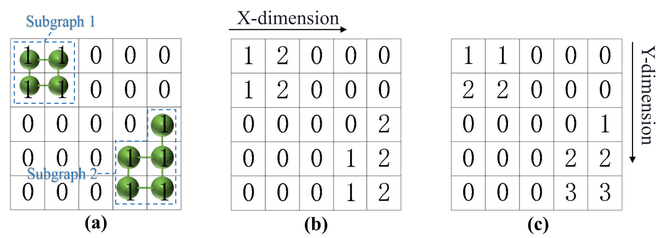


Fig. 7. Illustration of SPE. The subgraph mask (a) is first generated based on classification labels of trees; the subgraph represents regions with the same values in the presence of tree (1) or nontree (0). Then, the subgraph position in x dimension (b) and subgraph position in y dimension (c) are calculated using cumulative sum, and the calculation is reset when zero is encountered.

is first developed to obtain heterogeneous information. As the GCN layers become deeper, the node's receptive field expands, resulting in a smoothing effect. Therefore, the number of layers is set to 2. Each graph convolution is followed by batch normalization (BN) and activation functions. A residual block is added to improve the extraction of shallow features further. Thereby, high and low-frequency information can be obtained. The learned features are directly used for classification tasks (H_{class}) and simultaneously imported to SPEfor edge prediction (H_{link}) tasks. The SPE is designed to enhance the relative spatial position information of superpixel patches categorized as trees, thus strengthening the spatial position information between global graphs and compensating for the lack of natural spatial position information between two graphs. The detailed design will be further introduced in the next paragraph. After applying SPE, positional information is integrated into the output feature. In this process, the features that do not belong to the tree category have been removed, e.g., buildings surrounded by trees. The following equation describes the process of SGSFE:

$$\mathbf{H}_g^{l+1} = \mathrm{BN}\left(\mathrm{relu}\left(\widetilde{\mathbf{D}}_g^{-\frac{1}{2}}\widetilde{\mathbf{A}}_g\widetilde{\mathbf{D}}_g^{-\frac{1}{2}}\right)\mathbf{H}_g^l\Theta^l\right) \tag{3}$$

where BN represents batch normalization, l is the layer number, $\widetilde{\mathbf{A}}_g$ is the adjacency matrix $\mathbf{A}_g$ with self-loops added, $\widetilde{\mathbf{D}}_g$ is the diagonal degree matrix, and Θ^l is a trainable matrix in the lth layer. Especially, when $l = 0$, $\mathbf{H}_g^0 = \mathbf{X}_g$. In particular, the pertinent features of nodes are predominantly concentrated within the low-hop regions.

A detailed explanation of the SPE approach is provided here. The SPE was designed to explore spatial relationships between nodes by combining the use of subgraph masking and positional encoding. As graph structures already contain spatial information, they only possess adjacency information while lacking natural positional relationships between nodes. The natural spatial relationships among different pixels, superpixels, and subgraphs can be useful for both species classification and individual tree segmentation. In species classification, spatial texture details from superpixel blocks can be equivalent to spectral differences at the pixel level. Trees with distinctive crown textures, e.g., palm trees, tend to be classified into the same species. In individual tree segmentation, the spatial distribution of trees can assist in identifying crown edges. For example, roadside trees generally follow a specific pattern, and they can be easily separated based on their geographic locations. To perform SPE, a subgraph $\mathcal{G}_{tree}$ was first created based on the classification labels. Subgraphs represent collections of connected class-labeled tree superpixels, and there may be multiple subgraphs within a single graph, with each subgraph potentially containing multiple trees. Before generating the subgraphs, a subgraph mask is created based on the tree subgraphs, as seen in Fig. 7(a), where the values 1 and 0 indicate the presence and absence of trees, respectively. The regions composed of connected pixels with a value of 1 are referred to as subgraphs. For each subgraph, the corresponding bounding rectangle mask can be obtained from the subgraph mask. Then, the relative position indices of the bounding rectangle masks are calculated, and the cumulative sums are computed along specific dimensions (such as the x- and y-axes). Subsequently, the subgraph mask and relative position indices are multiplied element-wise to obtain the subgraph relative position indices. The calculation of the cumulative sums is illustrated in Fig. 7(b) and (c). Finally, a sine-based positional encoding technique utilized in the transformer [39] was adopted to calculate the relative position. The calculation for each element in the positional encoding matrix is shown in the following formula:

$$\begin{aligned}\mathrm{PE}_{(\mathrm{pos},2i)} &= \sin\left(\ \mathrm{pos}\ /10000^{2i/d_{\mathrm{model}}}\right)\\ \mathrm{PE}_{(\mathrm{pos},2i+1)} &= \cos\left(\ \mathrm{pos}\ /10000^{2i/d_{\mathrm{model}}}\right)\end{aligned} \tag{4}$$

where pos represents the positional index of a certain dimension within each subgraph relative to the bounding matrix of that subgraph, where i represents the dimension of the current positional encoding. Especially, d_{model} is the dimension of the outputs for SGSFE, which is set to be 128. Afterward, the SPE corresponding to each superpixel graph undergoes GAP to obtain the positional encoding for every node in the global graph. Subsequently, the positional encoding is combined with the feature $\mathbf{H}_{\mathrm{tree}}$ of the $\mathcal{G}_{\mathrm{tree}}$, resulting in the generation of the feature $\mathbf{H}_{\mathrm{link}}$.

B. MLP for Node Classification

After obtaining the features extracted by the SPM, the superpixel classification and superpixel aggregation results are obtained separately through an MLP classifier and a GCN decoder. The MLP is employed to predict the classification results $\mathbf{y_i}$ for each feature $\boldsymbol{h}_i^c$, which are utilized for node classification. This procedure can be mathematically represented as $\hat{y}_{\mathbf{i}} = \mathrm{MLP}(\boldsymbol{h}_i^c)$. Simultaneously, the node classification loss is measured by cross-entropy

$$L_{\mathrm{entropy}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \ln\left(\hat{y}_i\right). \tag{5}$$

C. GCN Decoder for Superpixel Aggregation

GCN decoder is employed to predict edges for superpixel aggregation, which is illustrated in Fig. 3, showing the design of the GCN decoder for edge prediction. First, the features $\mathbf{H}_l$ obtained through a scale pyramid are fed into the GCN layer. Subsequently, they undergo a nonlinear activation relu($\cdot$) and BN. Afterward, as shown in (6), we calculate the cosine similarity between each pair of neighboring nodes. Finally, the computed similarities undergo BN and sigmoid activation function to generate the prediction results $\hat{y}_{i,j}$

$$\text{Cosine Similarity}\left(\mathbf{h}_i^l, \mathbf{h}_j^l\right) = \frac{\mathbf{h}_i^l \cdot \mathbf{h}_j^l}{\|\mathbf{h}_i^l\|\|\mathbf{h}_j^l\|} \tag{6}$$

where $\mathbf{h}_i^l$ and $\mathbf{h}_j^l$ are feature vectors of pair of connected superpixels $\mathcal{A}_i$ and $\mathcal{A}_j$, respectively. Concurrently, the edge prediction loss is measured by binary cross entropy loss

$$L_{\mathrm{BCE}} = -\left[y_{i,j}\ln\left(\hat{y}_{i,j}\right) + \left(1 - y_{i,j}\right)\ln\left(1 - \hat{y}_{i,j}\right)\right]. \tag{7}$$

Thus, the final loss function is

$$L = L_{\mathrm{entropy}} + L_{\mathrm{BCE}}. \tag{8}$$

IV. Experiments and Results

This section describes the experiments of the SPGN method on UAV hyperspectral data for classifying individual tree species in a subtropical urban environment. In the evaluation of SPGN, metrics such as overall accuracy (OA), average accuracy (AA), and confusion matrix were calculated to assess the node classification performance of the model, while the area under the curve (AUC) was mainly used to measure the edge prediction performance. A series of ablation studies were conducted to identify the impacts of two key modules in SPGN, and method comparisons were also performed.

A. Datasets and Implementation Details

The UAV hyperspectral data were collected at the Yuehai Campus of Shenzhen University in Shenzhen, Guangdong, China. The true color image of the area is shown in Fig. 8, which was generated from three bands of hyperspectral data, where the band with a wavelength of 638 nm represents the red channel (R), the band with a wavelength of 552 nm represents the green channel (G), and the band with a wavelength of 472 nm represents the blue channel (B). The image was acquired in September 2022 using two DJI PHANTOM 4 RTK multirotor drones equipped with a Specim FX10 hyperspectral camera. The hyperspectral images comprise 112 spectral bands ranging from 400 to 1000 nm, with an average spectral resolution of 5.5 nm. The flight altitude was approximately 100 m above the ground, resulting in a spatial resolution of 10 cm for the hyperspectral images. A series of preprocessing steps were

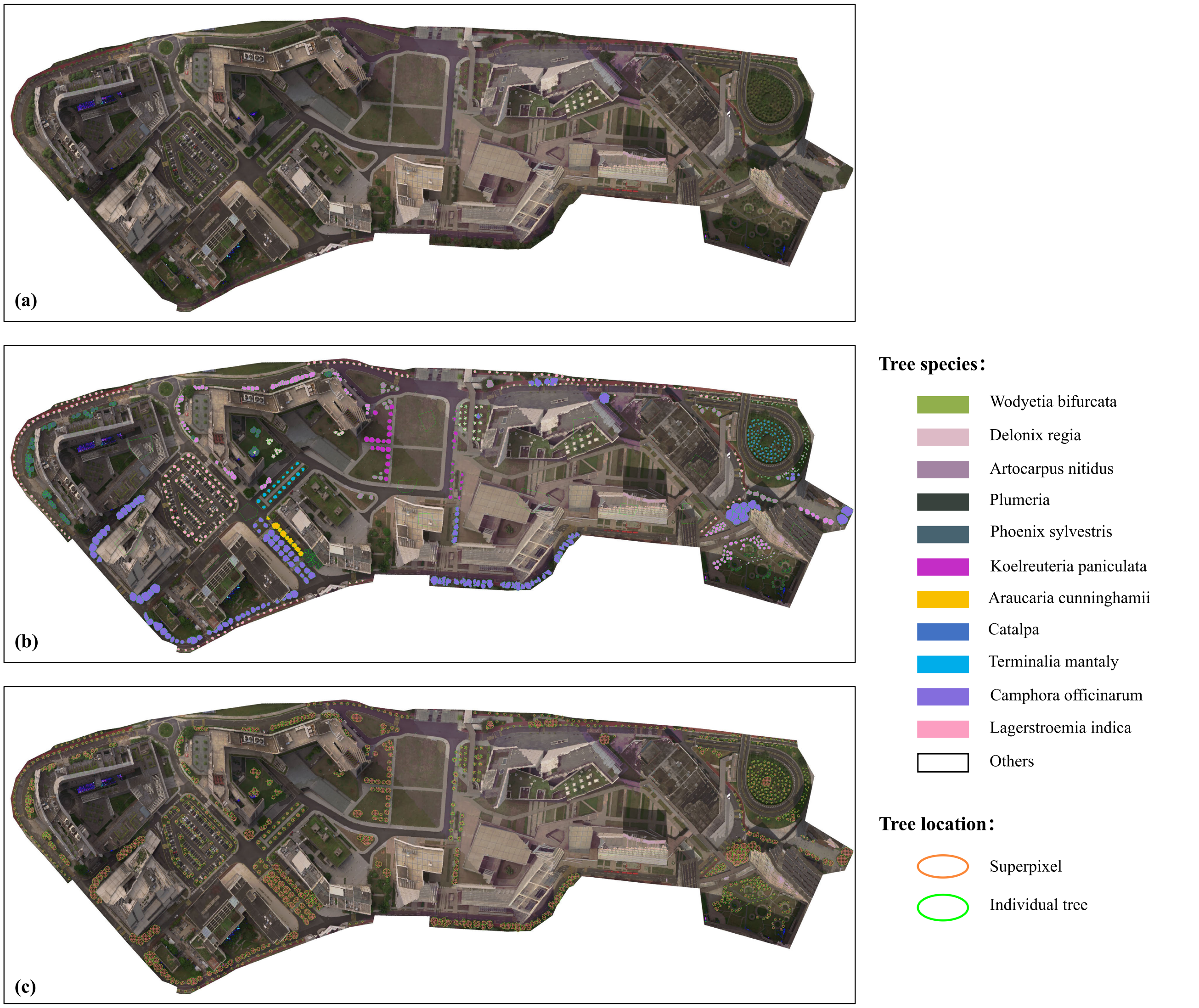


Fig. 8. Datasets and labeling of individual tree species in the study area. (a) True color representation (R: 638 nm, G: 552 nm, B: 472 nm) of the hyperspectral images of the study area, (b) manual labeling of trees at individual scales and corresponding species, where different colors represent different species, and (c) superpixel segmentation based on SLIC method.

implemented, including radiometric correction based on field reference data, geometric correction, atmospheric correction, spatial registration, color balancing, and image mosaic.

To obtain labeled samples for tree segmentation, field surveys on the tree species were conducted in the study area. Both the geo-locations of individual trees and the species names were identified [see Fig. 8(b)]. A total number of 835 trees belonging to 11 species were manually labeled, with the "others" category representing nontree objects. The species names and the specific number of trees in each category were documented in Table I. Spectra of labeled samples were averaged for each species, presenting a great similarity in spectral shapes as shown in Fig. 9(a). Moreover, within the same species, spectra from individual trees show small intravariability compared to that from different trees in Fig. 9(b), indicating the significance of combing species classification and tree segmentation.

For the node classification task, 20% of the superpixel nodes were used as training data, while 80% were used as testing data (see Table I). For the edge prediction task, 50% of the edges belonging to Link 0 and twice the number of Link 0 edges as Link 1 were used as training data, with the remaining edges used as the test set. The distribution of the segmented pixels is shown in Fig. 8. During the implementation of SPGN, the parameters were defined by several trials. To reduce resource consumption, K in (2) is set to 3, and the hidden units in all layers are set to 128. In addition, considering the size of individual tree pixel blocks, the value of k in SLIC is set to 100. For the training process, SPGN was trained using the Adam optimizer with a learning rate of 10^{-3} for 1000 epochs.

B. Tree Segmentation

Fig. 10 presents the tree segmentation results of the study area. Three representative regions were zoomed-in view to highlight details of segmentation and reveal more meaningful information. The SPGN-based segmentation was compared against superpixel segmentation results and ground truth

TABLE I
DATA DISTRIBUTION FOR EACH CATEGORY

NO.	Class	Abbreviation	Number of trees	Train (20%)	Test (80%)	All
C0	Others	OT		470	1,883	2,353
C1	Wodyetia bifurcata	WB	30	30	120	150
C2	Delonix regia	DR	69	82	331	413
C3	Artocarpus nitidus	AN	45	49	198	247
C4	Plumeria	PL	53	85	343	428
C5	Phoenix sylvestris	PS	33	74	300	374
C6	Koelreuteria paniculata	KP	9	24	97	121
C7	Araucaria cunninghamii	AC	101	49	198	247
C8	Catalpa	CA	18	21	86	107
C9	Terminalia mantaly	TM	127	493	1,973	2,466
C10	Camphora officinarum	CO	224	174	698	872
C11	Lagerstroemia indica	LI	126	158	636	794
Link 0	Belong to different trees	different		1,409 (50%)	1,409	2,818
Link 1	Belong to the same tree	same		2,818 (Twice as Link 0)	39,971	42,789

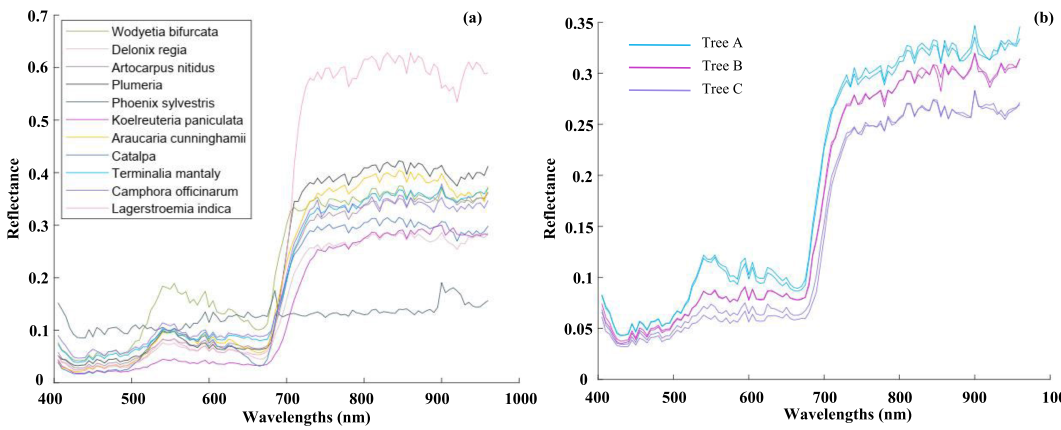


Fig. 9. Spectra of different species and different trees from UAV-HSI. (a) Average spectra of various tree species given in the study present significant differences in the spectral curves and (b) spectra from the patch of a single tree are more similar compared to those from different trees, within the same species.

in different scenarios: isolated and connected trees, regular shaped, and irregularly shaped trees. In all cases, SPGN successfully mapped tree crowns, which were mostly consistent with ground truth. In contrast, superpixel segmentation over-segmented a tree into several superpixel blocks (see Fig. 10). As shown in subfigures, SPGN achieved a good aggregation of superpixel blocks with similar features in the majority of samples. This indicates that SPGN has the potential to resolve the problem of over-segmentation. SPGN demonstrated the ability to distinguish connected trees with only a few minor defects. For instance, in region 1 of Fig. 10, connected trees (purple color) are well separated, although two trees are over-segmented. In addition, three trees close to the right side have also been identified as individuals, with some strange shapes on the edges. However, this odd crown delineation does not affect the final segmentation result; on the contrary, it proves that SPGN focuses on the edge features to distinguish overlapped trees, and the SPE module is very effective.

TABLE II
ACCURACY ASSESSMENT AND ABLATION STUDIES ON SPE AND SPM, MEASURED BY CLASSIFICATION ACCURACY, EDGE PREDICTION ACCURACY, AND AUC

Class	Abbrev.	SPGN-base	SPGN-SPE	SPGN-SPM	SPGN
C0	OT	97.24	96.12	99.36	**99.47**
C1	WB	65.83	51.67	**87.50**	85.83
C2	DR	73.11	62.54	**89.42**	83.68
C3	AN	77.27	65.15	91.92	**92.42**
C4	PL	83.97	78.43	**89.50**	86.88
C5	PS	84.33	84.00	94.00	**96.33**
C6	KP	84.54	58.76	**96.91**	95.88
C7	AC	91.92	90.90	**98.48**	**98.48**
C8	CA	82.56	68.60	95.35	**96.51**
C9	TM	93.72	90.87	**97.01**	96.15
C10	CO	82.66	81.94	90.26	**90.54**
C11	LI	83.33	77.35	92.92	**93.40**
AA (%)	–	83.37	75.53	93.55	**93.80**
OA (%)	–	89.42	85.71	95.42	92.97
Link 0	different	80.55	89.71	57.35	**92.97**
Link 1	same	83.52	86.53	**93.78**	90.27
AUC	–	0.89	0.93	0.89	**0.96**

C. Accuracy Assessment and Ablation Studies

Table II summarizes the classification and segmentation performance of SPGN, while the results of ablation studies highlight the impacts of different modules within SPGN. For species classification, the SPGN method achieved an AA of 93.80%. The species accuracy ranges from 83.68% (Delonix Regia) to 99.47% (Others). Most species can reach an accuracy

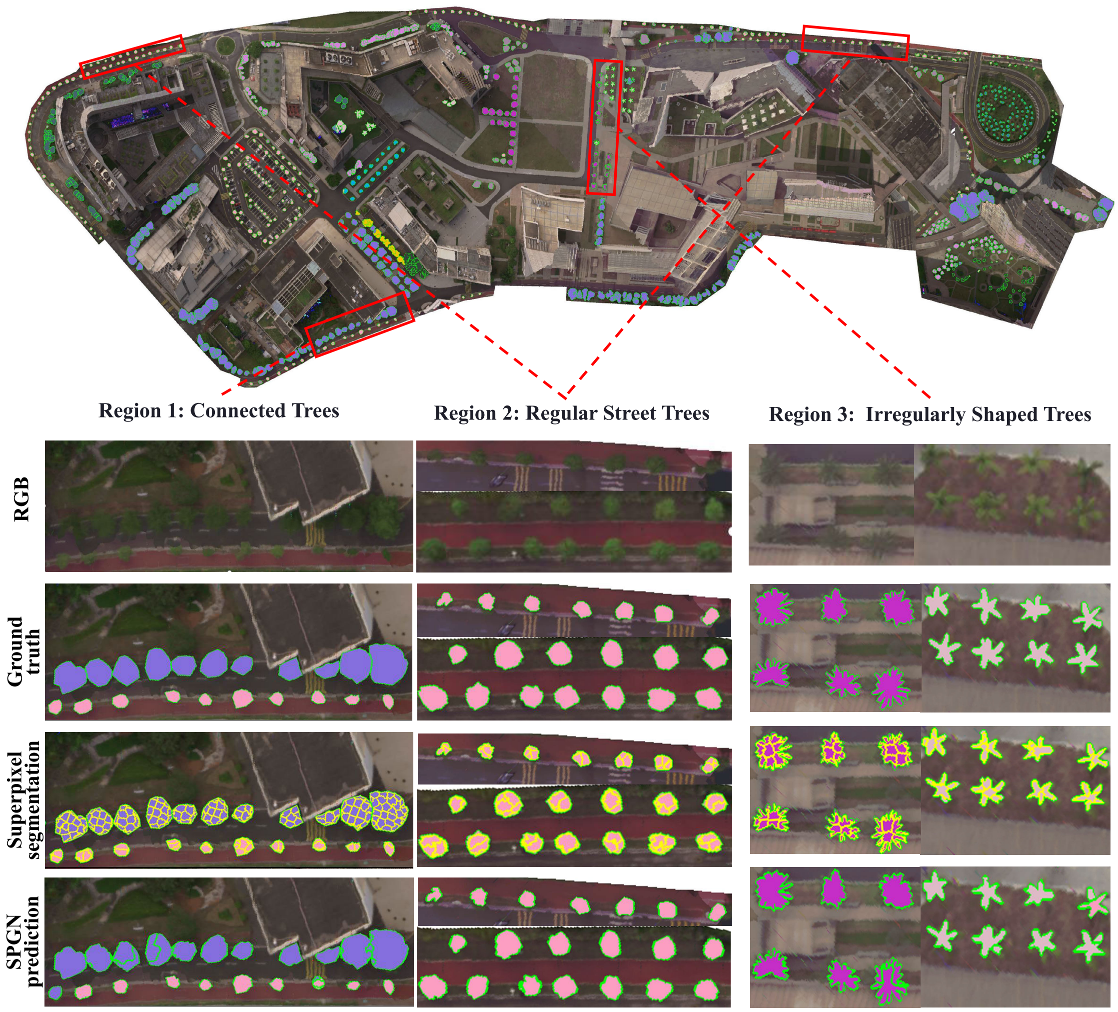


Fig. 10. Tree segmentation results in three zoomed-in view regions of connected trees, regular-shaped and irregular-shaped trees. SPGN prediction aggregated the over-segmented superpixels and agreed with the ground truth.

TABLE III
CONFUSION MATRIX OF SPECIES CLASSIFICATION USING SPGN

	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C0	1873	0	2	2	0	3	0	0	0	0	3	0
C1	0	103	0	0	0	0	2	2	0	2	9	2
C2	2	0	277	0	0	2	4	3	0	6	34	4
C3	0	0	0	183	0	1	3	0	0	6	2	3
C4	0	8	2	0	298	0	1	0	0	15	0	19
C5	2	2	0	0	0	289	2	0	0	1	3	1
C6	0	1	0	0	0	0	93	0	0	3	0	0
C7	0	0	0	0	0	0	0	195	0	0	2	1
C8	0	0	0	0	0	0	0	0	83	0	3	0
C9	9	12	5	9	0	0	1	1	0	1897	25	14
C10	4	2	19	1	4	0	0	0	3	17	632	16
C11	0	0	6	7	1	3	0	0	0	7	18	594

over 90%, indicating the capability of hyperspectral images for species classification. This accuracy is considerably high since all species belong to broadleaf plants and present great similarity in spectral signature. This can be attributed to the ultrahigh spatial and spectral resolution of UAV hyperspectral images. Table III presents the confusion matrix of SPGN-based classification. In the diagonal cells of the confusion matrix, a large number of correct predictions existed, indicating the high accuracy of SPGN for species classification. It also reveals some confusion between similar species, for example, DR and CO, since the DR species had no florals during the time of data collection. CO is also easily confused with TM, both species present dense canopy and evergreen broad leaves.

From Table II, it can be observed that the SPM greatly improved the classification, as evidenced by the highest OA of 95.42%. It increased the OA by 6% compared to the original network without SPM (SPGN-base). For individual tree segmentation, SPGN derived an accuracy of 92.97% for link 0 (adjacent superpixels belonging to different trees) and 90.97% for link 1 (same tree). This is a significant improvement compared to the accuracy of 80.55% and 80.52% from the original model. The improvement made by the SPE module is more significant, which is plausible because this module was designed to enhance the edge prediction performance. The increase in segmentation accuracy can be further demonstrated by the AUC number. SPGN with SPE improved the original AUC from 0.89 to 0.93. Therefore, we generally speculate that SPM improved the feature extraction of the network because it can extract different levels of feature representation through the scale pyramid mechanism, while SPE focuses on identifying individual trees by merging similar superpixel blocks, capturing natural spatial relations through positional encoding techniques.

The combined use of SPM and SPE together improves both classification and segmentation, as reflected by the increased OA, link, and AUC metrics. However, their specific impact on the final performance is unpredictable. For example, the SPM does not function well in segmenting individual trees, particularly for overlapped trees (only 57.35% accuracy for link 0 cases). This might be attributed to the following reasons. Incorporating the scale pyramid without positional encoding would make the network biased in extracting features for classifying species, thus neglecting the edges of connected trees. Therefore, superpixels belonging to the different trees can hardly be separated. On the other hand, the SPM does not affect the segmentation of isolated trees, as represented by the 93.78% for link 1, and it actually improves their segmentation. This can be because the features of isolated trees are already representative enough, and the scale pyramid feature extraction promotes the merging of adjacent pixels. Conversely, the addition of SPE decreases the classification accuracy from 89.42% to 85.71%, which is also reasonable given that more attention is directed toward edge prediction rather than feature extraction for species discrimination. The separation of individual trees relies more on subtle changes in tree crown shapes and small texture differences between connected regions of adjacent trees. The addition of two modules further increases the accuracy to 92%, probably because SPM improves public feature learning ability, and this public feature can accommodate classification and segmentation tasks at the same time. Therefore, it is advisable to use these two compatible modules, SPE and SPM, together.

To further examine the impact of different modules for individual tree species classification, we present the data distribution using t-distributed stochastic neighbor embedding (t-SNE), through which the separability between different species can be visualized. Fig. 11 shows the sample distribution maps after feature extraction by SPGN-base, SPGN-SPE, SPGN-SPM, and SPGN. It can be seen from Fig. 11(a) and (b) that the sample distribution maps in the 11 tree species based on methods without SPM have a high degree of overlap after being projected into the 2-D space, including SPGN-base and SPGN-SPE. Comparing the two subfigures of Fig. 11(a) and (c), it can be found that the features based on methods without the SPM are not discriminative, so the sample distributions of different classes have a large overlap. Based on methods with the SPM, samples between different classes are well separated and samples within the same class

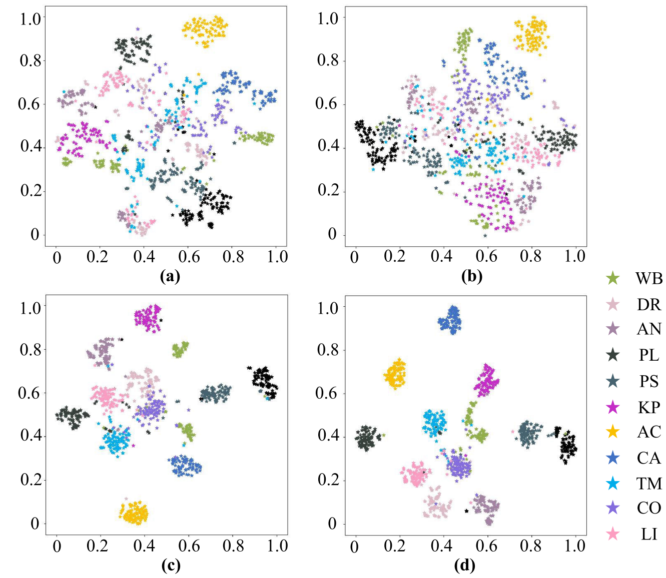


Fig. 11. Class separation using t-SNE, based on (a) SPGN-base, (b) SPGN-SPE, (c) SPGN-SPM, and (d) SPGN.

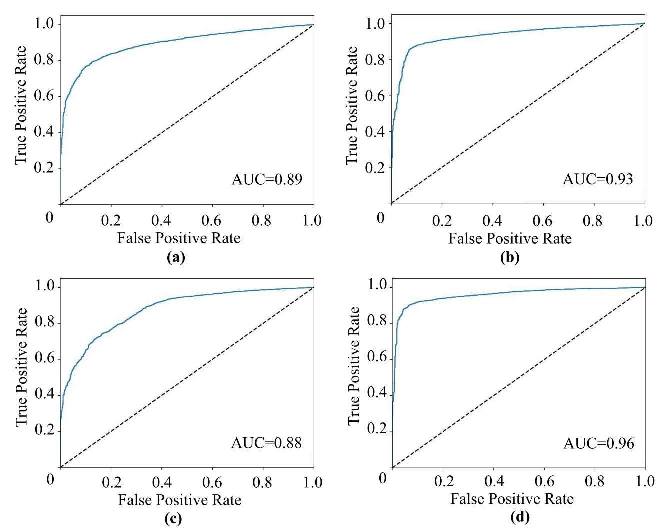


Fig. 12. ROC for ablation studies: SPGN provides the best performance, indicating the advantage of SPE. (a) SPGN-base. (b) SPGN-SPE. (c) SPGN-SPM. (d) SPGN.

are clustered, due to more discriminative features. The most obvious improvement is the separation of the KP, AC, and CA. Similar to the quantitative results in ablation studies (see Table II), the SPE module did not improve the classification of these spectrally similar targets, whereas the SPM greatly enhanced the separation of all classes, as reflected by the increased intervariability and decreased intravariability from SPGN-SPM and SPGN which includes both SPE and SPM. Samples with similar features are clustered, indicating the importance of extracting multiscale features for the species classification using a graph network. Plots from SPGN show more compact clusters, which indicates the SPE also has a positive effect on the classification. It can be hypothesized that knowing the positional information of superpixels may assist their species identification.

The positive impact of SPE on tree segmentation can be further demonstrated by the receiver operating curve (ROC) curve

Fig. 13. Method comparison for individual tree segmentation at selected area. (a) RGB image of the region, (b) results of the DeepForest method, where the red bounding boxes indicate the detected individual trees, (c) results of the Detectree2 method, where different colors represent different individual trees, (d) segmentation results of SAM, where different colors represent different instance objects, (e) segmentation results of SPGN, where each block represents an individual tree and different colors represent their respective categories, and (f) ground truth.

of the ablation studies in Fig. 12. A ROC curve helps evaluate how well a binary classifier distinguishes between classes (tree versus nontree) by plotting the true positive rate (TPR) on the vertical axis against the false positive rate (FPR) on the horizontal axis. The closer this curve is to the top-left corner of the graph, the better the classifier's performance. Comparing ROC curves from different methods, SPGN exhibits the best performance. SPGN has a higher AUC (0.96) than SPGN-base (0.89), highlighting the advantage of SPE. Both SPGN-SPE and SPGN outperform SPGN-base in edge prediction tasks. This indicates that positional encoding effectively enhances the spatial structure of the graph, thereby improving the model's ability to extract spatial texture information. SPE only encodes the positions for adjacent superpixel blocks rather than constructing positional encoding relative to the entire graph, resulting in more accurate positional encoding. There is a slight decrease in AUC when using SPM without SPE, from 0.89 to 0.88. This can be ascribed to the same reason as previously explained in the ablation studies.

D. Method Comparison

Fig. 13 shows a comparison of the SPGN method with three other methods: DeepForest, Detectree2, and segment anything model (SAM) [48]. As described in the related work section, DeepForest is a pretrained network based on the RetinaNet model [49] designed to simplify training models for tree detection. To clarify, DeepForest and SAM methods can only be applied to RGB images; therefore, three-band images representing red (650 nm), green (550 nm), and blue (450 nm) channels were spectrally resampled from HSI. The SPGN identified individual trees in the selected mixed tree and urban environment areas, while all other methods missed the identification of some trees. DeepForest failed to detect trees, and Detectree2 only detected one tree and misidentified a road as a tree. SAM was able to detect several tree individuals with perfect boundaries but still missed some samples. This might be attributed to the shadow effect in this particular area. Due to the lack of annotated RGB image data, we did not fine-tune the comparison models and only made predictions on RGB images from locally collected regions. SPGN detected all trees but over-segmented one tree. The species classification made by SPGN agreed well with the ground truth, with

TABLE IV
RESULTS (%) OF DIFFERENT INSTANCE SEGMENTATION NETWORK MODELS

Method	Mask mAP_{50} (%)	Parameters
Mask R-CNN	44.50	43.80 M
SS-FPN Mask R-CNN	79.20	47.34 M
SS-FPN SOLOv1	65.40	39.46 M
SS-FPN SOLOv2	72.20	49.59 M
SPGN Mask R-CNN	**83.17**	44.42 M

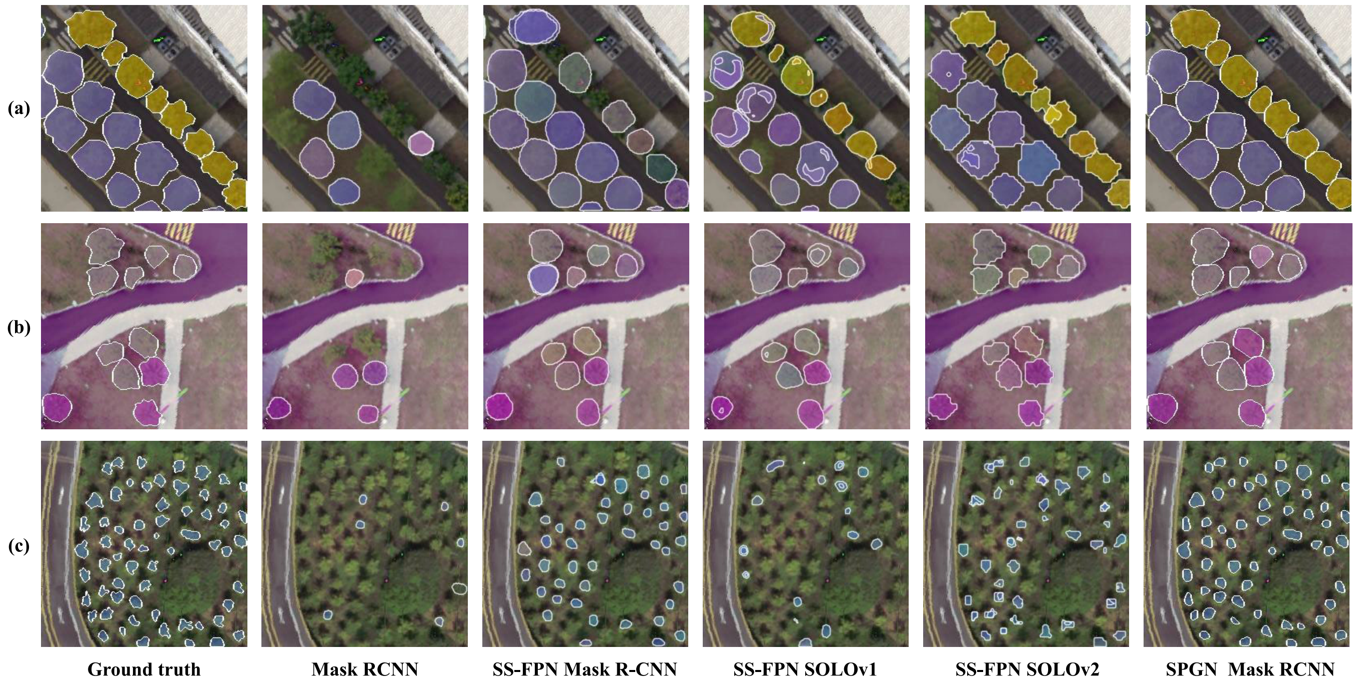


Fig. 14. Method comparison for individual tree segmentation by SPGN mask R-CNN, mask R-CNN, SS-FPN mask R-CNN, SS-FPN SOLOv1, and SS-FPN SOLOv2 in three representative regions (a) overlapping and intersecting trees, (b) discrete street trees, and (c) dense saplings.

only one misclassified tree. The remaining samples were correctly identified as the Lagerstroemia indica species. This level of accuracy is considered very high, given that only 20% of training data was used. This figure is only a typical representation of the study area, whereas the rest of the areas generally follow this pattern.

To further demonstrate the effectiveness of the SPGN method, this article introduces mask R-CNN, SS-FPN mask R-CNN, SS-FPN SOLOv1, and SS-FPN SOLOv2 as comparative methods. Among them, mask R-CNN is a representative instance segmentation method in the RGB image domain. In this article, hyperspectral data is dimensionally reduced to three channels through principal component analysis (PCA) and used as the input for the model. SS-FPN mask R-CNN, SS-FPN SOLOv1, and SS-FPN SOLOv2 are the first instance segmentation models tailored for hyperspectral images, and hyperspectral data is directly employed as the input for these models. To rationalize the experimental setup for comparisons, the core module of SPGN, namely, SPM, is integrated into mask R-CNN (termed SPGN mask R-CNN), unifying the experimental conditions across all methods. Considering that trees and other ground objects can be readily distinguished using NDVI and classifiers, the locations of trees are directly obtained from labels in this experiment, and corresponding subgraph position encodings are generated as inputs to the models. The original data are cropped into 512 $\times$ 512 images, with 80% randomly selected as training data and the remaining 20% as testing data.

As shown in Table IV, SPGN mask R-CNN attains the highest mAP_{50}, reaching 83.17%, which represents a 38.67% improvement over the baseline method, mask R-CNN, and is 3.97% higher than SS-FPN mask R-CNN. This result sufficiently demonstrates that the core module of the SPGN method, SPM, can effectively extract features from hyperspectral images, exhibiting a stronger capability for spatial-spectral feature extraction. Furthermore, SPGN mask R-CNN has a parameter count of only 44.42M, a mere 0.62M more than mask R-CNN, indicating an extremely low parameter volume. This parameter count is significantly lower than that of SS-FPN mask R-CNN, achieving better results with fewer parameters.

As evident from Fig. 14, the mask R-CNN method is able to segment a limited number of trees but suffers from significant missed detections, resulting in decreased accuracy in detected tree categories. In contrast, SS-FPN mask R-CNN can detect the majority of trees, albeit with a relatively high misclassification rate for tree categories. It demonstrates a relatively better segmentation ability for smaller trees but still encounters numerous missed detections. SS-FPN SOLOv1 excels in recognizing larger trees but experiences severe missed detections for smaller trees. However, its ability to distinguish tree categories is superior to SS-FPN Mask R-CNN. SS-FPN SOLOv2, while competent in recognizing both large and small trees with some improvement for the latter, exhibits a slight decline in its capability to distinguish tree categories compared to SS-FPN SOLOv1. Notably, SPGN mask R-CNN not only achieves remarkable detection for larger trees but also closely aligns with ground truth for smaller trees. Although there are some misdetections across all categories, its overall performance is the closest to the ground truth. Therefore, it can be concluded that SPGN mask R-CNN achieves the best performance among the compared methods in terms of visual prediction results, further validating the strong feature extraction capability of the SPM module.

V. Conclusion

This article proposes a new paradigm of instance segmentation specifically for trees from ultrahigh spatial resolution hyperspectral images. An SPGN was designed by combining the task of species classification and edge prediction for tree segmentation. This graph-based network works well on extracting features for irregular objects (e.g., tree crowns) and overcomes the over-segmentation by using traditional superpixel methods. A high accuracy of 93% and AUC of 0.96 was achieved based on UAV hyperspectral images of pedestrian broadleaf trees in a subtropical urban environment (University campus) in South China, proving the effectiveness of SPGN. The ablation studies of the method demonstrate the positive impacts of two modules, SPM and SPE, in species classification and tree segmentation. SPM extracts hyperspectral image features at the pixel, superpixel, and subgraph scales, thereby better exploring the potentially useful information for differentiating species at each scale. SPE overcomes the limitation of graph structure only containing adjacency information and partially compensates for the lack of natural spatial positional information, thus improving the model's detection capability for edge prediction tasks. The combination of two modules within SPGN maximizes the potential of high spectral and spatial resolution in UAV hyperspectral images, offering an affordable and convenient way for individual tree mapping. Future research may focus on applying and adapting the SPGN method to identify a broader range of tree species across diverse ecosystems and geographical settings.

References

[1] L. Chang, H. Fan, N. Zhu, and Z. Dong, "A two-stage approach for individual tree segmentation from TLS point clouds," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8682–8693, 2022.

[2] J. G. C. Ball et al., "Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using mask R-CNN," *Remote Sens. Ecol. Conservation*, vol. 9, no. 5, pp. 641–655, Oct. 2023.

[3] R. Engler et al., "Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution," *Forest Ecol. Manage.*, vol. 310, pp. 64–73, Dec. 2013.

[4] R. Mcroberts and E. Tomppo, "Remote sensing support for national forest inventories," *Remote Sens. Environ.*, vol. 110, no. 4, pp. 412–419, Oct. 2007.

[5] L. Deng et al., "Comparison of 2D and 3D vegetation species mapping in three natural scenarios using UAV-LiDAR point clouds and improved deep learning methods," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103588.

[6] B. Fu et al., "Combination of super-resolution reconstruction and SGA-Net for Marsh vegetation mapping using multi-resolution multispectral and hyperspectral images," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2724–2761, Oct. 2023.

[7] M. Dalponte and D. A. Coomes, "Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data," *Methods Ecol. Evol.*, vol. 7, no. 10, pp. 1236–1245, Oct. 2016.

[8] M. F. Gomes, P. Maillard, and H. Deng, "Individual tree crown detection in sub-meter satellite imagery using marked point processes and a geometrical-optical model," *Remote Sens. Environ.*, vol. 211, pp. 184–195, Jun. 2018.

[9] G. Lassalle, M. P. Ferreira, L. E. C. La Rosa, and C. R. de Souza Filho, "Deep learning-based individual tree crown delineation in mangrove forests using very-high-resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 220–235, Jul. 2022.

[10] L. Wang, W. P. Sousa, P. Gong, and G. S. Biging, "Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of Panama," *Remote Sens. Environ.*, vol. 91, nos. 3–4, pp. 432–440, Jun. 2004.

[11] C. Tucker et al., "Sub-continental-scale carbon stocks of individual trees in African drylands," *Nature*, vol. 615, no. 7950, pp. 80–86, Mar. 2023.

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

[13] B. G. Weinstein, S. Marconi, M. Aubry-Kientz, G. Vincent, H. Senyondo, and E. P. White, "DeepForest: A Python package for RGB deep learning tree crown delineation," *Methods Ecol. Evol.*, vol. 11, no. 12, pp. 1743–1751, Dec. 2020.

[14] Y. Gan, Q. Wang, and A. Iio, "Tree crown detection and delineation in a temperate deciduous forest from UAV RGB imagery using deep learning approaches: Effects of spatial resolution and species characteristics," *Remote Sens.*, vol. 15, no. 3, p. 778, Jan. 2023.

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

[16] I. Marin, S. Gotovac, and V. Papic, "Individual olive tree detection in RGB images," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2022, pp. 1–6.

[17] G. P. Asner, T. R. Seastedt, and A. R. Townsend, "The decoupling of terrestrial carbon and nitrogen cycles," *BioScience*, vol. 47, no. 4, pp. 226–234, Apr. 1997.

[18] G. P. Asner, "Biophysical and biochemical sources of variability in canopy reflectance," *Remote Sens. Environ.*, vol. 64, no. 3, pp. 234–253, Jun. 1998.

[19] Y. Long, B. Rivard, A. Sanchez-Azofeifa, R. Greiner, D. Harrison, and S. Jia, "Identification of spectral features in the longwave infrared (LWIR) spectra of leaves for the discrimination of tropical dry forest tree species," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 97, May 2021, Art. no. 102286.

[20] F. E. Fassnacht et al., "Review of studies on tree species classification from remotely sensed data," *Remote Sens. Environ.*, vol. 186, pp. 64–87, Dec. 2016.

[21] S. K. Meerdink, D. A. Roberts, K. L. Roth, J. Y. King, P. D. Gader, and A. Koltunov, "Classifying California plant species temporally using airborne hyperspectral imagery," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111308.

[22] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[23] G. T. Miyoshi et al., "A novel deep learning method to identify single tree species in UAV-based hyperspectral images," *Remote Sens.*, vol. 12, no. 8, p. 1294, Apr. 2020.

[24] H. Qin, W. Zhou, Y. Yao, and W. Wang, "Individual tree segmentation and tree species classification in subtropical broadleaf forests using UAV-based LiDAR, hyperspectral, and ultrahigh-resolution RGB data," *Remote Sens. Environ.*, vol. 280, Oct. 2022, Art. no. 113143.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[26] F. Schiefer et al., "Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 205–215, Dec. 2020.

[27] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *Int. J. Multimedia Inf. Retr.*, vol. 9, no. 3, pp. 171–189, Sep. 2020.

[28] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[29] X. Xi, K. Xia, Y. Yang, X. Du, and H. Feng, "Evaluation of dimensionality reduction methods for individual tree crown delineation using instance segmentation network and UAV multispectral imagery in urban forest," *Comput. Electron. Agricult.*, vol. 191, Dec. 2021, Art. no. 106506.

[30] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 23, 2024, doi: 10.1109/TNNLS.2023.3336563.

[31] T. Zhang et al., "Semantic attention and scale complementary network for instance segmentation in remote sensing images," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10999–11013, Oct. 2022.

[32] L. Fang, Y. Jiang, Y. Yan, J. Yue, and Y. Deng, "Hyperspectral image instance segmentation using spectral–spatial feature pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5502613.

[33] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 115–134, Apr. 2019.

[34] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[38] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Pro. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 21–37.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. Conf. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[40] K. Chowdhary and K. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*. Cham, Switzerland: Springer, 2020, pp. 603–649.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[42] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2017, pp. 1–14.

[44] F. Wu, A. H. Souza, T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6861–6871.

[45] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.

[46] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 1157–1171, Jan. 2024.

[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[48] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

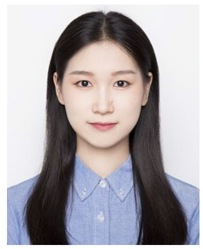
[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

Yaqian Long received the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2019.

She is currently an Assistant Professor with Shenzhen University, Shenzhen, China. Her study area is hyperspectral remote sensing in geological and ecological applications. Her primary research topic focuses on feature selection, classification, and fusion using spectral mixture analysis and state-of-the-art machine-learning techniques.

Songxin Ye received the B.S. degree in network engineering from Foshan University, Foshan, China, in 2021. He is currently pursuing the master's degree in computer technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and deep learning.



Liqiong Wang received the B.S. degree in computer science and technology from Henan University of Economics and Law, Zhengzhou, China, in 2023. She is pursuing the master's degree in computer technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include hyperspectral image unmixing.

Weixi Wang received the Ph.D. degree in geodesy and survey engineering degree from Liaoning Project Technology University of China, Huludao, China, in 2007.

He was a Post-Doctoral Fellowship with Wuhan University, Wuhan, China, in 2013. He is currently an Associate Professor with Shenzhen University, Shenzhen, China. He is mainly engaged in research on building information modeling (BIM), 3-D geographical information systems (GISs), and virtual geographic environments (VGEs).

Xiaomei Liao received the B.S. degree in atmospheric science from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in physical oceanography from South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, in 2016.

She is currently an Associate Researcher with the College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. Her research interests include tropical ocean dynamical processes and deep learning.

Sen Jia (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include hyperspectral image processing, signal and image processing, and machine learning.