Texture-Aware Self-Attention Model for Hyperspectral Tree Species Classification

Nanying Li, Shuguo Jiang, Jiaqi Xue, Songxin Ye, and Sen Jia, *Senior Member, IEEE*

***Abstract*— Forests play an irreplaceable role in carbon sinks. However, there are obvious differences in the carbon sink capacity of different tree species, so the scientific and accurate identification of surface forest vegetation is the key to achieving the double carbon goal. Due to the disordered distribution of trees, varied crown geometry, and high difficulty in labeling tree species, traditional methods have a poor ability to represent complex spatial–spectral structures. Therefore, how to quickly and accurately obtain key and subtle features of tree species to finely identify tree species is an urgent problem to be solved in current research. To address these issues, a texture-aware self-attention model (TASAM) is proposed to improve spatial contrast and overcome spectral variance, achieving accurate classification of tree species hyperspectral images (HSIs). In our model, a nested spatial pyramid module is first constructed to accurately extract the multiview and multiscale features that highlight the distinction between tree species and surrounding backgrounds. In addition, a cross-spectral–spatial attention module is designed, which can capture spatial–spectral joint features over the entire image domain. The Gabor feature is introduced as an auxiliary function to guide self-attention to autonomously focus on latent space texture features, further extract more appropriate and accurate information, and enhance the distinction between the target and the background. Verification experiments on three tree species hyperspectral datasets prove that the proposed method can obtain finer and more accurate tree species classification under the condition of limited labeled samples. This method can effectively solve the problem of tree species classification in complex forest structures and can meet the application requirements of tree species diversity monitoring, forestry resource investigation, and forestry carbon sink analysis based on HSIs.**

***Index Terms*— Carbon sinks, hyperspectral images (HSIs), spectral–spatial attention module, tree species classification.**

I. Introduction

CARBON peaking and carbon neutrality have become the focus of global attention and also our country's national strategy. As the country's urbanization process continues to advance, effectively increasing carbon sinks is the key to carbon peaking and carbon neutrality strategies [1], [2], [3]. Carbon sink mainly refers to the amount of carbon dioxide absorbed and stored by forests, or the ability of forests to absorb and store carbon dioxide. In other words, forests play an irreplaceable role in carbon sinks [4], [5]. Due to the obvious differences in carbon sink capacity among different tree species, scientifically and accurately identifying surface forest vegetation and analyzing forestry carbon sinks are the key to achieving carbon peak and carbon neutral goals.

Traditional tree species identification mainly relies on professional foresters to carry out a field investigation and identify tree species based on external morphological characteristics such as roots, trunks, branches, leaves, flowers, and fruits [6], [7], [8]. Although this method is relatively accurate, it also has many shortcomings, and it is difficult to meet the needs of high-precision tree species information extraction. First of all, limited by the complex environment, poor accessibility, and other factors, it is impossible to obtain comprehensive and detailed information [9]. Second, manual field surveys are costly and time-consuming, so it is difficult to achieve macroscale tree species identification in a short time [10]. Finally, this method has high requirements on the professional level of the staff and requires them to have a deep tree species knowledge reserve. Moreover, the different cognitive standards of each person will also lead to errors in the work, and the final results will be rough.

In recent years, with the continuous development of remote-sensing technology, the acquisition of high spatial resolution images is more convenient and faster. The application of remote-sensing technology is becoming more and more extensive [11], which provides the possibility for accurate and rapid tree species classification [12], [13], [14]. Hyperspectral images (HSIs) have broken through the bottleneck that makes it difficult to capture the subtle spectral differences of tree species in natural images and are a key technical means to achieve tree species classification [15], [16]. For example, Gong [17] utilized the approximate nearest neighbor (ANN) classification method to discriminate the spectral data, thereby distinguishing one broad-leaved tree species from six coniferous tree species and obtaining higher classification accuracy. Martin et al. [18] excavated the relationship between AVIRIS hyperspectral data and the chemical composition of tree species leaves and identified eleven tree species types, which can indeed effectively classify tree species. Subsequently, Koedsin and Vaiphasa [19] demonstrated that EO-1

Manuscript received 21 June 2023; revised 13 September 2023 and 29 October 2023; accepted 4 December 2023. Date of publication 19 December 2023; date of current version 3 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290; and in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, the Guangdong–Hong Kong–Macau Joint Laboratory for Smart Cities, and the Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, Ministry of Natural Resources, Shenzhen University, Shenzhen 518060, China (e-mail: linanying2021@email.szu.edu.cn; shuguoj@foxmail.com; xuejiaqi2021@email.szu.edu.cn; yesongxin2021@email.szu.edu.cn; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3344787

spaceborne hyperspectral data can be used to distinguish tropical mangrove species. They applied a classical genetic search algorithm to reduce the dimension of HSIs and a spectral angle mapping (SAM) classifier to classify them. Harrison et al. [20] collected leaf reflectance spectra of twenty-six tree species in the long-wave infrared region and observed that most of the spectral features were derived from cell walls or cuticle compounds. Then, the author used wavelet transform to preprocess the spectrum and used random forest to classify tree species. In the same year, Hycza et al. [21] took temperate forests in the Walmia Nature Reserve in Poland as the research object and used a supervised classification method to classify seven dominant tree species in the HSIs. This experiment obtained a high accuracy rate, which also proved that HSIs can accurately classify tree species in natural forests. In addition, Fricker et al. [22] found that HSIs have obvious advantages over ordinary visible light images in tree species identification.

Although HSIs can provide rich spectral information, the original spectral curves of various tree species canopies are very similar. Only relying on the original spectral features to classify tree species is easily affected by the fact that different species have the same spectrum, thereby reducing the classification performance. Therefore, adding features such as texture features, vegetation indices, and mathematical statistics features can effectively improve the classification accuracy [23]. Zhang et al. [24] converted the spatial–spectral features extracted from the original data into 1-D features and then used them as new inputs for a 3-D convolutional neural network (3D-CNN) model. This method improves the identification accuracy of tree species in the case of scattered, unclear boundaries, and canopy occlusion. Tong and Zhang [25] proposed a spectral–spatial and cascaded multilayer random forests (SSCMRFs) method, which concatenates the output of superpixel-based classification and spectral features, and then integrates them into spatial information. Moreover, Guo et al. [26] simulated the multisource information fusion process and input the morphological and spectral features of tree species into a dual-concentrated network with morphological features (DNMFs). This method decouples spatial and spectral information and obtains better classification results. Next, Zhang et al. [27] proposed a novel spatial logical aggregation network (SLA-NET) to effectively integrate spectral, spatial, and detailed morphological information of tree species in HSIs. This framework focuses on fine-grained morphological structural characteristics of tree species and finds the boundaries of fine-grained categories through spatial morphological differences. However, for areas with complex forest structure conditions, dense tree crowns, and numerous species, no method can achieve fine tree species classification.

With the introduction and development of precision forestry and digital forestry, forest management requires higher and higher spatial and spectral resolutions of images. Some researchers have combined light detection and ranging (LiDAR) data and hyperspectral data to classify forest tree species and have made positive progress. For instance, Liu et al. [28] combined the spectral features of hyperspectral data and the canopy height features of LiDAR to effectively distinguish tree species with similar spectra but different heights. Zhao et al. [29] combined LiDAR data and HSI data to classify natural mixed forests in northeast China. First, the LiDAR data and HSI data are registered. Then, the improved watershed algorithm is used to segment the LiDAR data into a single tree, and the single tree segmentation area is mapped to the hyperspectral data to extract the single tree spectrum. Finally, support vector machines (SVMs) and SAM are employed to classify tree species. Next year, Mäyrä et al. [30] collected hyperspectral data, LiDAR data, and extensive ground-referenced data for the research area in the northern cold zone of southern Finland. The ground-based reference data were then matched to the aerial imagery by a canopy-level model derived from LiDAR. Finally, the performance of different traditional methods such as 3D-CNN, random forest, and SVM was compared, and it was found that 3D-CNN can more effectively distinguish conifer species and have higher classification accuracy. However, the acquisition cost of LiDAR data is expensive and is affected by the area of the flight area, so the application prospect is limited [31], [32]. The rapid development of near-earth low-altitude unmanned aerial vehicle (UAV) remote-sensing platforms provides a new opportunity for fine tree species classification and identification based on hyperspectral data [33]. The portability, economy, and speed of the UAV platform provide great convenience for forestry workers. Workers can quickly acquire images even in specific research areas and harsh mountainous areas. While reducing the workload, it can also obtain higher quality data, which greatly reduces the difficulty of forest resource investigation. UAVs equipped with hyperspectral remote sensors can obtain HSIs with submeter spatial resolution through low-altitude imaging. It provides fine texture information and continuous spectral information for tree species classification, making vegetation identification more detailed and accurate.

Therefore, we use UAVs equipped with hyperspectral remote sensors to obtain tree species hyperspectral data and propose a texture-aware self-attention model (TASAM) to sufficiently mine multiview and -scale features and spatial–spectral features of tree species. It consists of a nested pyramid module and cross-spectral–spatial attention (a stack of alternative spectral-wise self-attention and Gabor-guided spatial-context self-attention). The model exploits views from near to far and scales from small to big, providing distinct spatial contexts and textures of targets for the subsequent module. Meanwhile, it takes into consideration spectral correlation on the whole image domain. Thus, our texture-aware can overcome low spatial contrast and huge spectral variance on tree species HSIs. The shallow and Gabor-guided architecture design enables the model to perform well and accurately identify various tree species even with few labeled samples.

Concretely, it takes a patch pyramid as input and performs a series of scale-aware operators on each layer to generate corresponding scale maps. Since scale maps progressively decrease from top to bottom, the resultant is a nest spatial pyramid, where the outer part is a patch pyramid while the inner part of each layer is an image one. It integrates various spatial context and scale features of targets under the nested

structure to improve their contrast. Next, the multiview and multiscale features are carried into the cross-spectral–spatial attention module. Here, spectral-context self-attention takes spectral-wise tokens as input and calculates their correlation via self-attention. This mechanism considers spectral correlation under the whole image domain instead of performing an affine transformation on one pixel. Consequently, self-attention is very suitable for overcoming huge spectral variances of tree species at developing stages. As known to all, naive self-attention is insensitive to the locality and transformation variance on images, so it is required to learn in massive data, which serves as image prior knowledge. However, this demand is not satisfied in the field since annotating images is expensive. In addition, considering the subtle spectral differences caused by factors such as water and chlorophyll in different growth and development stages of forest trees, Gabor-guided spatial-context self-attention is designed. This module can effectively combine spectral features and texture features to improve the classification accuracy of tree species, and at the same time, it can also enhance the distinction between the phenomenon of "different species with the same spectrum" and "same species with different spectra." In the end, a simple fully connected layer is used to predict the final classification result of pixels. The overall architecture of texture-aware self-attention is shown in Fig. 1. The main contributions of this article can be summarized as follows.

1) First, a nested spatial pyramid module is proposed to extract multiview and multiscale features of targets against their low spatial contrast in tree species HSIs. Due to the influence of many factors such as water vapor in the field environment, sun intensity, and complex background, the spectral differences between different tree species will become smaller. Meanwhile, they are also hard to discriminate from the context in some cases. All of these easily bewilder models and incur misclassification results, which motivates us to improve their contrast. To this end, the nested spatial module takes as input an image pyramid containing multiple different receptive fields and performs a series of scale-aware operators for each layer to produce the corresponding feature pyramid. The resultant present a nested pyramid structure, where the outer part is the image pyramid and the inner part of each layer is the feature one. This highly integrated spatial structure contains spatial–spectral difference information from surrounding neighboring pixels as well as multiscale features concerning targets to produce the final representation with discriminability and robustness.
2) Considering that factors such as water and chlorophyll in different growth and development stages of tree species will cause subtle changes in spectral information, we also design a cross-spectral–spatial attention module to capture correlation within spectra on the whole image domain and Gabor-guided spatial texture features alternatively. In essence, cross-spectral–spatial attention is a stack of alternative spectral-context self-attention and Gabor-guided spatial-context self-attention. Unlike current popular schemes that regard pixels or patches as tokens, our spectral-context self-attention converts each band map as one token and takes the spectral-wise sequence of the whole image as input. Then, self-attention dynamically calculates their correlation and updates them. In addition, each band map incorporates specific spatial characteristics under different wavelengths, but self-attention is not well-designed to capture them. Thus, we introduce Gabor features as an auxiliary role in guiding self-attention to focus on latent spatial characteristics autonomously, capturing the correlation between targets.
3) According to standards such as coverage type, planned utilization route, and canopy density, woodland types can be divided into ten first-level land types such as arbor forest land, mangrove land, sparse forest land, and nursery land. Even so, the proposed model is suitable for different types of forest land and can perform fine classification of tree species. Besides, due to its shallow architecture and combination of hand-crafted features, our network still performs well, even with a few labeled samples. Extensive ablation and comparison experiments on three different types of woodland datasets confirm that our model can overcome low spatial contrast and huge spectral variance for different tree species and achieve finer HSI tree species classification than state-of-the-art networks.

The remainder of this article is organized as follows. Section II briefly introduces the related work. Section III introduces the proposed method in detail. Afterward, Section IV elaborates on the experimental settings and results systematically. Finally, in Section V, we draw some conclusions.

II. Related Work

A. Transformer

Transformer [34], [35] is first proposed in natural language processing (NLP) to replace recurrent neural networks (RNNs) [36], [37] for performing language translation since it can capture long-range dependencies and run in parallel. Its core mechanism, self-attention, is shown as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}}{\sqrt{d_k}}\right)\mathbf{V}. \tag{1}$$

The attention takes queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ as input and measures the similarity between $\mathbf{Q}$ and $\mathbf{K}$ to update elements $\mathbf{V}$ of the sequence. The vector dimensions in $\mathbf{Q}$ and $\mathbf{K}$ are both d_k. The mechanism calculates weights dynamically according to current inputs so that the transformer can be seen as a general modeling method but depends on massive labeled data to train. Over recent years, the transformer has shown a powerful ability in modeling sequence relationships, attracting much attention from computer vision and stimulating a round of revolution there.

Vision transformer (ViT) [38] is the first entirely self-attention-based architecture proposed for image classification, where the image is split as a series of 16×16 tokens along spatial dimensions and passes through the naive transformer network to get the final representation. This successful

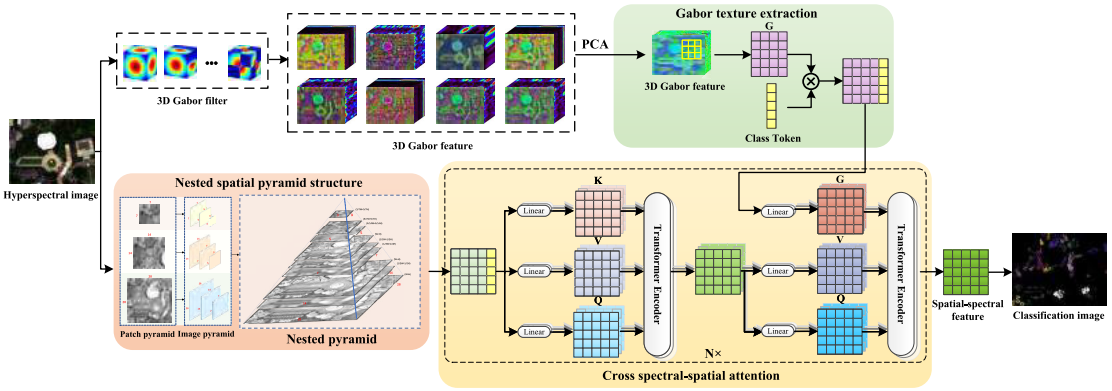

Fig. 1. Overall framework of the proposed TASAM for HSI tree species classification.

application of the transformer to image classification demonstrates convolutional neural networks (CNNs) [39], [40] is not the unique option and breaks its long-time domination in computer vision. Compared with CNNs, the architecture of the transformer is more efficient and skilled in capturing long-short-range dependencies among objects. Nevertheless, it initially is not designed to process images and constantly works on low-resolution inputs during processing without considering hierarchy. Moreover, locality is ignored there, so a great deal of data is required to learn. To this end, Wang et al. [41] propose a pyramid structure to reduce spatial resolution progressively and generate different partitions hierarchically. Thus, the model receives low-resolution inputs at shallow layers and high-resolution ones at deep layers. In addition, Liu et al. [42] designed a shifting window mechanism that restricts self-attention in predefined windows and shifts the windows along spatial dimensions to perform information interaction. This approach improves running efficiency and preserves long-range extraction for self-attention. Most importantly, it introduces the locality as prior knowledge about images to make the model better applicable for computer vision tasks. Given the effectiveness and flexibility of the transformer in image classification, it has subsequently been extended to other computer vision tasks, such as HSI classification, to produce better results. For instance, Qing et al. [43] proposed an end-to-end transformer model based on the spectral attention module and the multihead self-attention module. Unlike CNNs, this model uses fewer convolutions to achieve better classification performance. Subsequently, Tu et al. [44] introduced the concept of homogeneous regions into ViT and proposed a framework based on a local semantic feature aggregation transformer (LSFAT). This framework mainly consists of feature aggregation and self-attention calculation, which provides good classification results and acceptable computing performance on HSI classification.

Nevertheless, the transformer highly depends on massive labeled samples as prior knowledge to learn a more representative, semantic, and robust feature space. To alleviate this requirement, many efforts attempt to fuse the transformer with CNNs, exploiting long-range extraction from the former and local spatial perception from the latter. Gao et al. [45] embed the self-attention into CNNs to enlarge its receptive field, and Zhang et al. [46] integrated both of them in a parallel manner to capture features from global and local levels simultaneously. However, they all put too much attention on pixel-wise interaction and only use a simple feed-forward network to perform spectral transformation on single pixels. Consequently, existing methods based on the transformer may focus on spatially coarse features and neglect spectrally detailed ones when faced with HSIs where spatial resolution is much smaller than spectral resolution. On the other hand, the fusion of the transformer and CNNs to enhance spatial description increases training load indispensably, which is unfavorable when the number of training samples is limited. So, we should consider how to leverage abundant spectral and spatial characteristics via the transformer here.

TABLE I

TREE SPECIES ON THE YUEHAI DATASET AND THE NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR EACH TREE SPECIES

Class	Name	Train	Test
1	Ficus microcarpa	324	32080
2	Ficus altissima Blume	1094	108339
3	Litchi chinensis	541	53620
4	Dimocarpus longan	287	28496
5	Araucaria cunninghamii	264	26159
6	Pinus	150	14888
7	Cinnamomum camphora	134	13325
8	Ficus elastica	678	67171
9	Livistona chinensis	57	5736
10	Leucaena leucocephala	48	4828
11	Roystonea regia	143	14249
12	Mangifera indica	71	7036
13	Terminalia arjuna	914	90493
14	Delonix regia	97	9618
15	Kigelia africana	62	6200
16	Syzygium cumini	151	14957
	Total	5015	497195

B. Three-Dimensional Gabor

In image processing, the Gabor function [47], [48] is a linear filter used for edge extraction. The frequency and direction expressions of Gabor filters are similar to those of the human visual system, and they can provide good direction-selective and scale-selective properties [49]. Additionally, Gabor filters are insensitive to lighting transformations, as they are well

TABLE II

TREE SPECIES ON THE CANGHAI DATASET AND THE NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR EACH TREE SPECIES

Class	Name	Train	Test
1	Cinnamomum camphora	205	20343
2	Terminalia neotaliala	341	33786
3	Ceiba insignis	86	8590
4	Plumeria rubra	68	6804
5	Roystonea regia	15	1516
6	Dracontomelon duperreanum	37	3727
7	Mangifera indica	25	2566
8	Koelreuteria paniculata	42	4198
9	Artocarpus parvus	54	5352
10	Kigelia africana	43	4291
11	Lagerstroemia indica	123	12243
	Total	1039	103416

TABLE III

TREE SPECIES ON THE ZHANJIANG DATASET AND THE NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR EACH TREE SPECIES

Class	Name	Train	Test
1	Rhizophora stylosa	249	24743
2	Talipariti tiliaceum	95	9458
3	Excoecaria agallocha	126	12503
4	Bruguiera gymnorhiza	580	57455
5	Kandelia obovata	90	8973
6	Aegiceras corniculatum	1013	100314
7	Sonneratia apetala	821	81378
	Total	2974	294824

TABLE IV

ABLATION ANALYSIS OF TASAM WITH DIFFERENT FUSION MANNERS

Dataset	Spectral self-attention	Image pyramid	Patch pyramid	Texture attention	OA(%)	AA(%)	Kappa
Yuehai	✓	✗	✗	✗	69.07	48.18	0.64
	✓	✓	✗	✗	81.69	70.80	0.79
	✓	✗	✓	✗	84.87	79.23	0.83
	✓	✓	✓	✗	88.74	82.19	0.87
	✓	✓	✓	✓	**89.83**	**84.49**	**0.88**
Canghai	✓	✗	✗	✗	82.44	76.91	0.78
	✓	✓	✗	✗	86.16	82.25	0.83
	✓	✗	✓	✗	86.08	79.77	0.83
	✓	✓	✓	✗	90.74	88.93	0.89
	✓	✓	✓	✓	**92.15**	**90.72**	**0.90**
Zhanjiang	✓	✗	✗	✗	80.88	66.04	0.74
	✓	✓	✗	✗	84.31	73.08	0.79
	✓	✗	✓	✗	85.18	74.79	0.80
	✓	✓	✓	✗	86.10	75.95	0.82
	✓	✓	✓	✓	**89.03**	**82.71**	**0.86**

suited for texture representation and separation. The 2-D Gabor function [50], [51] is the convolution of the 2-D Gaussian function and 2-D Fourier function. The 2-D Gabor function can be expressed as

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = e^{-\frac{x'^2+y^2y^2}{2\sigma^2}} e^{i\left(2\pi \frac{x'}{\lambda}+\psi\right)} \quad (2)$$

$$x' = x\cos\theta + y\sin\theta \quad (3)$$

$$y' = -x\sin\theta + y\cos\theta \quad (4)$$

where x and y are the position coordinates. k defines the Gabor scale. λ is the wavelength, which directly affects the filtering scale of the filter. θ specifies the direction of the parallel fringes of the Gabor function. ψ is the phase offset. σ represents the standard deviation of the Gaussian factor in the Gabor function. γ determines the shape of the filter.

In the spatial domain, a 2-D Gabor kernel is actually the result of a Gaussian kernel modulated by a sine wave. The usual form of the Gabor filter kernel function is

$$G_g(k, x, y, \theta) = \frac{k^2}{\sigma^2} e^{-\frac{k^2\left(x^2+y^2\right)}{2\sigma^2}} \left(e^{ik(x\cos\theta+y\sin\theta)} - e^{-\frac{\sigma^2}{2}}\right) \quad (5)$$

where σ is the standard deviation of the Gaussian function. A Gabor kernel function can obtain the response of a frequency neighborhood of the image, and the response result can be regarded as a feature of the image. If multiple Gabor kernels with different frequencies are used to obtain the response of images in different frequency neighborhoods, the characteristics of images in each frequency band can be formed, which can describe the frequency information of images. Since texture features are usually associated with frequency, the Gabor kernel is often used to extract texture features. Based on this, by exploiting phase-induced Gabor kernels, Liu et al. [52] proposed a Gabor-nets that can automatically adapt to the local harmonic features of HSI data and can generate more representative harmonic features.

The 3-D Gabor filter [53] can find a good combination of localization in the frequency and spatial domains. Therefore, it can take into account the spectral characteristics of the image based on extracting texture information. Jia et al. [54] proposed a collaborative representation-based multiscale superpixel fusion (CRMSF) method for HSI classification, which applies 3-D Gabor filters to extended multiattribute profiles (EMAPs) features to represent the internal spatial–spectral structure of HSIs. Subsequently, a flexible Gabor-based superpixel-level unsupervised LDA method for HSI classification was proposed [55]. This method introduces the degree of freedom control parameters to modify the generalized sinusoidal plane wave in the kernel Gabor wavelet, thus accurately improving the expressiveness of capturing various structural distributions of materials.

III. METHODOLOGY

In the face of low spatial contrast and high spectral variance of targets, existing methods are hard to perform finer classification for tree species HSIs. Moreover, they also work poorly when only a few labeled samples are provided there. To overcome these shortages, we propose a texture-aware transformer that consists of a nest spatial pyramid module and a cross spectral–spatial attention module (a stack of alternative spectral-context self-attention and Gabor-guided spatial-context self-attention) for spectral–spatial feature extraction. Specifically, it can mine multiview and multiscale features from inputs, aggregate spectral characteristics on the whole image domain, and exploit latent spatial patterns guided by Gabor. Concurrently, the well-design shallow architecture and combination of hand-crafted features enable our model still accurately identify tree species. As a result, the model presents accurate classification under the case with low spatial contrast, high spectral variance, and limited samples. Next, we will describe the model's overall pipeline and some key designs in detail. The overall flowchart of the proposed method is shown in Fig. 1.

First, given an HSI $\mathbf{I} \in \mathbb{R}^{H\times W\times B}$, where H, W, and B denote its height, width, and band number, respectively, there

TABLE V

CLASSIFICATION PERFORMANCE BY USING THE SEVEN COMPARED METHODS FOR THE YUEHAI HYPERSPECTRAL DATASET WITH 1% LABELED SAMPLES PER CLASS AS TRAINING SET (NUMBERS IN BOLD REPRESENT THE BEST CLASSIFICATION PERFORMANCE)

Class	Training	SVM	CNN	3D-1D-CNN	SpectralFormer	SCAT	SSFTT	SSCL3DNN	TASAM
1	324	23.63 ± 15.27	37.12 ± 11.47	**74.36 ± 3.12**	31.43 ± 10.22	69.12 ± 6.89	67.32 ± 6.83	57.20 ± 2.25	72.51 ± 7.47
2	1094	62.07 ± 15.86	76.63 ± 4.69	91.15 ± 1.36	78.44 ± 7.56	87.09 ± 5.50	82.43 ± 6.98	81.38 ± 1.47	**91.20 ± 2.01**
3	541	28.40 ± 15.69	70.18 ± 9.09	86.06 ± 2.24	75.04 ± 8.43	85.21 ± 2.83	78.79 ± 7.60	80.78 ± 1.54	**86.09 ± 5.31**
4	287	23.02 ± 18.53	44.38 ± 15.96	**82.89 ± 3.86**	49.19 ± 14.84	66.70 ± 8.29	63.85 ± 14.21	61.96 ± 4.50	77.37 ± 4.05
5	264	89.02 ± 6.92	93.86 ± 7.65	98.58 ± 0.49	97.18 ± 1.29	97.82 ± 0.69	96.94 ± 1.58	90.33 ± 1.15	**99.04 ± 0.38**
6	150	5.56 ± 11.07	38.57 ± 11.36	86.77 ± 5.43	49.14 ± 23.86	84.39 ± 4.12	78.46 ± 5.98	52.09 ± 3.63	**92.11 ± 4.63**
7	134	18.18 ± 25.47	14.83 ± 11.57	**77.68 ± 4.68**	20.20 ± 8.52	65.20 ± 11.04	56.74 ± 5.67	44.63 ± 4.32	75.31 ± 7.20
8	678	83.82 ± 6.53	98.30 ± 1.14	95.39 ± 2.26	96.79 ± 2.45	98.98 ± 0.33	98.78 ± 0.87	96.62 ± 0.57	**99.40 ± 0.64**
9	57	17.95 ± 13.26	40.93 ± 16.75	**82.76 ± 6.45**	43.06 ±15.33	74.40 ± 7.71	62.48 ± 9.29	58.89 ± 5.46	72.60 ± 14.02
10	48	17.60 ± 16.49	30.77 ± 24.53	**77.61 ± 8.81**	38.65 ± 23.96	78.57 ± 16.08	52.43 ±10.04	46.35 ± 7.71	75.82 ± 8.73
11	143	15.15 ± 14.27	26.34 ± 11.90	73.06 ± 8.82	40.68 ± 13.53	63.80 ± 7.45	52.32 ± 7.37	42.02 ± 2.87	**73.32 ± 6.61**
12	71	5.27 ± 4.32	42.33 ± 21.91	80.59 ± 1.97	36.99 ±21.87	82.76 ± 7.51	66.92 ± 11.09	77.03 ± 4.26	**89.75 ± 5.45**
13	914	80.28 ± 6.94	92.41 ± 5.28	95.85 ± 1.54	89.13 ± 3.70	95.19 ± 1.09	95.32 ± 0.98	88.99 ± 0.97	**98.28 ± 0.87**
14	97	15.32 ± 20.77	20.03 ± 15.48	71.25 ± 5.53	48.20 ± 19.31	77.97 ± 9.72	63.58 ± 8.33	71.60 ± 5.73	**81.71 ± 4.98**
15	62	11.13 ± 18.48	0.53 ± 1.45	52.51 ± 9.20	54.75 ± 31.15	63.90 ± 9.12	48.33 ± 4.71	42.82 ± 8.77	**70.00 ± 9.41**
16	151	43.19 ± 28.61	81.15 ± 5.40	96.02 ± 1.19	73.09 ± 12.34	93.20 ± 2.90	85.94 ± 5.30	93.01 ± 1.52	**97.39 ± 1.33**
OA		33.72 ± 2.10	50.52 ± 4.13	82.66 ±2.69	57.62 ± 9.85	80.27 ± 2.86	71.92 ± 3.49	67.86 ± 1.35	**84.49 ± 1.90**
AA		53.25 ± 2.07	70.65 ± 2.08	88.71 ± 1.69	72.74 ± 5.17	86.24 ± 2.14	82.14 ± 3.83	78.55 ± 0.70	**89.83 ± 1.61**
Kappa		0.46 ± 0.02	0.66 ± 0.03	0.87 ± 0.02	0.68 ± 0.06	0.84 ± 0.02	0.80 ± 0.04	0.75 ± 0.01	**0.88 ± 0.02**

are many varied-scale views reflecting global contexts and local details of tree species in multiple scales. Those views $\mathbf{X}_\alpha \in \mathbb{R}^{\alpha h \times \alpha w \times B}$ centered on the same target pixel from small to big form the image pyramid, which is input to the nested spatial pyramid module in the front of the texture-aware transformer. Here, h and w are the smallest height and width of the views, and α is the magnification. In the module, a sequence of M scale-aware operators is performed on the image pyramid and generates corresponding scale maps as an inner feature pyramid for each layer. Note that each view incorporates B maps so $\mathbf{X}_\alpha = [\mathbf{X}_\alpha^1, \ldots, \mathbf{X}_\alpha^b, \ldots, \mathbf{X}_\alpha^B]$. Considering that they record and reveal distinct spatial details concerning tree species at one specific wavelength, the scale-aware operators process them separately. Thus, the resultant scale maps for each is $\{\mathbf{S}_{\alpha,\beta}^b\}_{\beta=1}^M$. Following the general setup, they from 1 to M gradually decrease a half spatial scale. Then, a feed-forward network (FFN) takes these maps lying in different views as input and outputs a multiview and multiscale feature $\mathbf{Z}_i$ that presents spatial details with high differences concerning tree species at the ith band. As a result, the final representation $\mathbf{Z} = [\mathbf{Z}^1, \ldots, \mathbf{Z}^B] \in \mathbb{R}^{h \times w \times B}$. Next, $\mathbf{Z}$ is transmitted into the cross-spectral–spatial attention module to refine spectral and spatial features further. In this module, the spectral-context self-attention builds spectral correlations and captures salient ones on the whole image domain, which is helpful for our model to overcome huge spectral variances of tree species as they grow. Besides, Gabor-guided spatial-context self-attention further digs hidden patterns in each band through hand-crafted features to enhance spatial details.

A. Nested Spatial Pyramid Module

One single view has limited valuable information, especially in the case that the tree species is similar to surrounding pixels due to spectral changes caused by shadow, complex backgrounds, and so on. In addition, crowns are not presented on a fixed scale since their scale and shape vary as they grow. Consequently, we design a nested spatial pyramid, where the outer is a patch pyramid and the inner is image one, to show various view-level and scale-level features. The structure is generated by our proposed nested spatial pyramid module, which takes a patch pyramid constructed from patches of different sizes as input, and performs a series of adaptive mean pooling $f(\cdot)$ on each patch, illustrated in Fig. 1. The processing is shown as follows:

$$\mathbf{S}_{\alpha,\beta}^b = f_{\frac{\alpha h}{2^{\beta-1}} \times \frac{\alpha w}{2^{\beta-1}}}\left(\mathbf{X}_\alpha^b\right). \tag{6}$$

Here, the footnote $(\alpha h/2^{\beta-1}) \times (\alpha w/2^{\beta-1})$ represents the output scale. Afterward, the scale maps $\{\mathbf{S}_{\alpha,\beta}^b\}_{\beta=1}^M$ are stacked together from small to big and from top to bottom as an image pyramid. Subsequently, the image pyramid of each patch is rearranged into a nested spatial pyramid structure.

To combine multiview and multiscale features for subsequent processing, FFN processing is performed on the nested spatial pyramid to obtain multiview and multiscale feature representations with the same spatial size, shown as follows:

$$\mathbf{Z}^b = \mathrm{FFN}\left(\left[\mathbf{S}_{1,1}^b, \ldots, \mathbf{S}_{\alpha,\beta}^b, \ldots\right]\right). \tag{7}$$

Note that each map would be padded with zero to keep the same size. In the end, the final feature representation is $\mathbf{Z} = [\mathbf{Z}^1, \ldots, \mathbf{Z}^B]$, which is transmitted into the cross-spectral–spatial attention.

B. Cross-Spectral–Spatial Attention Module

After the nested spatial pyramid module, our texture-aware transformer is followed by a cross-spectral–spatial attention module. It is composed of N blocks, each of which stacks a piece of spectral-context self-attention and Gabor-guided spatial-context self-attention to alternatively refine spectral and spatial features. Next, we will discuss their design in detail.

1) Spectral-Context Self-Attention: Besides presenting spatially geometry structures, HSIs also yield spectral curves that reflect chemistry concerning tree species. Hyperspectral remote sensing can obtain more detailed vegetation ecological information by estimating the content of biophysical and chemical components of different types of vegetation.

TABLE VI

CLASSIFICATION PERFORMANCE BY USING THE SEVEN COMPARED METHODS FOR THE CANGHAI HYPERSPECTRAL DATASET WITH 1% LABELED SAMPLES PER CLASS AS THE TRAINING SET (NUMBERS IN BOLD REPRESENT THE BEST CLASSIFICATION PERFORMANCE)

Class	Training	SVM	CNN	3D-1D-CNN	SpectralFormer	SCAT	SSFTT	SSCL3DNN	TASAM
1	205	79.49 ± 22.47	86.29 ± 7.76	90.03 ± 4.75	89.35 ± 4.18	92.99 ± 3.05	94.44 ± 1.48	85.61 ± 1.00	**95.35 ± 1.38**
2	341	81.95 ± 10.49	82.97 ± 10.03	95.48 ± 3.11	**96.15 ± 2.30**	95.55 ± 1.77	92.82 ± 6.00	92.17 ± 0.90	95.47 ± 1.60
3	86	38.43 ± 25.61	70.36 ± 11.19	81.61 ± 8.51	82.42 ± 7.14	84.73 ± 5.19	82.82 ± 7.07	84.25 ± 4.32	**89.53 ± 4.09**
4	68	48.44 ± 35.27	37.18 ± 21.54	77.09 ± 13.81	86.29 ± 3.30	**89.47 ± 5.21**	75.15 ± 30.16	78.82 ± 4.09	85.46 ± 8.44
5	15	17.61 ± 27.70	51.94 ± 13.69	70.67 ± 8.03	72.21 ± 15.27	80.46 ± 13.68	82.78 ± 9.47	63.09 ± 9.72	**90.53 ± 4.69**
6	37	54.86 ± 14.05	68.05 ± 6.90	79.55 ± 14.71	77.26 ± 5.95	**91.05 ± 5.53**	84.85 ± 9.50	80.02 ± 4.04	86.94 ± 3.22
7	25	37.11 ± 40.50	90.33 ± 21.34	86.35 ± 13.30	92.97 ± 3.91	96.94 ± 3.09	97.13 ± 2.75	96.57 ± 1.78	**97.97 ± 1.81**
8	42	33.77 ± 22.32	30.27 ± 21.91	75.14 ± 17.66	57.98 ± 10.49	**79.61 ± 5.63**	72.78 ± 21.63	73.88 ± 2.72	77.56 ± 6.45
9	54	54.71 ± 27.61	38.39 ± 22.43	83.73 ± 10.89	89.64 ± 7.84	89.07 ± 6.01	90.48 ± 3.70	78.91 ± 3.92	**93.16 ± 2.21**
10	43	19.37 ± 20.56	78.56 ± 20.58	80.45 ± 15.79	97.80 ± 1.80	89.69 ± 9.83	96.63 ± 2.83	88.51 ±2.95	**97.15 ± 1.50**
11	123	48.39 ± 26.45	69.24 ± 13.98	80.10 ± 8.23	81.88 ± 5.61	86.39 ± 4.29	81.76 ± 9.13	**90.66 ± 1.89**	88.79 ± 3.76
OA		46.74 ± 8.54	63.96 ± 5.22	81.84 ± 10.22	83.99 ± 2.55	88.72 ± 4.60	86.51 ± 8.35	82.95 ± 2.33	**90.72 ± 3.14**
AA		62.32 ± 7.16	72.07 ± 3.58	86.70 ± 6.60	87.89 ± 1.93	90.82 ± 3.04	88.38 ± 6.88	86.50 ± 0.68	**92.15 ± 1.94**
Kappa		0.53 ± 0.09	0.65 ± 0.05	0.84 ± 0.08	0.85 ± 0.02	0.89 ± 0.04	0.86 ± 0.08	0.84 ± 0.01	**0.90 ± 0.02**

Green plants have obvious spectral reflection characteristics, that is, the spectral reflection or emission characteristics of vegetation are determined by their chemical composition and morphological characteristics, which are closely related to the development, health, and growth conditions of vegetation. Specifically, in the visible light band, various pigments are the main factors that dominate the spectral response of plants, among which chlorophyll plays the most important role. In the near-infrared band of the spectrum, the spectral properties of vegetation are mainly controlled by the internal structure of plant leaves. In the midinfrared band of the spectrum, the spectral response of green plants is dominated by the strong absorption band of water. However, due to factors such as water vapor in the field environment, solar intensity, and complex background, as well as the changes in water and chlorophyll caused by the growth and development of the tree species itself, the spectral curves of different tree species will also be different.

Therefore, faced with such complex spectral changes in tree species, previous models that rely on an FFN to perform spectral/channel transformation on one single pixel never consider their correlation on the whole image domain. They put too much attention on extracting spatial features and underestimate abundant spectral context. As a result, those are hard to adapt to huge spectral variance when identifying tree species HSIs and to be transferred to other morphology and individual images. To address this problem, we propose spectral-context self-attention that converts HSIs into spectral-wise sequences instead of spatial-wise ones and dynamically captures spectral relationships on the whole image domain. First, $\mathbf{Z}$ is reshaped as $\widetilde{\mathbf{Z}} \in \mathbb{R}^{hw \times B}$, where each token in the sequence is a band map. In the article, our final goal is to obtain pixel-wise classification results, so a class token $\boldsymbol{c}$ is appended behind the sequence and serves as its representation at the end, where $\widehat{\mathbf{Z}} = [z_1, \ldots, z_B, \boldsymbol{c}]$.

Then, spectral-context self-attention receives the sequence and extracts globally spectral context features in tree species HSIs. Its concrete operation is shown as follows:

$$\widetilde{\mathbf{H}}_l = \mathbf{V}_l \cdot \text{softmax}\left(\frac{\mathbf{Q}_l^T \mathbf{K}_l}{\sqrt{d_k}}\right) \tag{8}$$

where

$$\mathbf{Q}_l = \mathbf{W}_l^{\mathbf{Q}} \mathbf{H}_{l-1}, \quad \mathbf{K}_l = \mathbf{W}_l^{\mathbf{K}} \mathbf{H}_{l-1}, \quad \mathbf{V}_l = \mathbf{W}_l^{\mathbf{V}} \mathbf{H}_{l-1}. \tag{9}$$

And l refers to the layer number of the encoder in our self-attention structure. During processing, $\mathbf{H}_{l-1}$ coming from the previous layer is converted into the query $\mathbf{Q}_l$, key $\mathbf{K}_l$, and value $\mathbf{V}_l$ through three linear mappings $\mathbf{W}_l^{\mathbf{Q}}$, $\mathbf{W}_l^{\mathbf{K}}$, and $\mathbf{W}_l^{\mathbf{V}}$. In particular, $\mathbf{H}_0 = \widehat{\mathbf{Z}}$. The attention mechanism calculates the correlation between spectral-wise tokens by multiplying transposed $\mathbf{Q}_l$ and $\mathbf{K}_l$ matrices and normalizes their numerical values by softmax$(\cdot)$. It aggregates highly similar features from $\mathbf{V}_l$, according to the calculated attention matrix. Note that our aggregation is along the spectral dimension on the whole image instead of along the spatial one where attention is softmax$(((\mathbf{Q}_l \mathbf{K}_l^T)/\sqrt{d_k}))\mathbf{V}_l$. In this way, our attention matrix that is generated dynamically can adapt to spectral variances in different developing stages for tree species as well as distinguish them well from surrounding similar pixels. Furthermore, it takes into consideration global spectral correlation on the whole image domain, overcoming huge spectral differences and producing robust representations.

2) Gabor-Guided Spatial-Context Self-Attention: Although the nested pyramid module in the front of our model reveals multiview and multiscale features concerning tree species, flattening each band map into a token above is not beneficial for the model to capture spatial context further. On the other hand, as known to us, self-attention is insensitive to spatial details and needs much data to learn, which is hard to meet here.

Inspired by information retrieval, self-attention can be interpreted as its matching process. Here, $\mathbf{Q}$ is users' queries, and $\mathbf{K}$ is the key words associated with specific objects. As a result, the elements of the attention matrix $\mathbf{Q}\mathbf{K}^T$ reflect matching degrees for each other. While $\mathbf{K}$ in the naive self-attention is learned by a linear transformation, it depends on much-labeled data and does not have very explicit spatial geometry meanings. Gabor features extracted by a series of hand-crafted filters in various orientations and scales can characterize inherent image textures and do not introduce extra trainable parameters. Considering its excellent representing ability for images and low demand for labeled data, we design

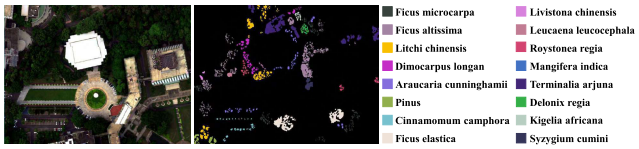

Fig. 2. False-color map and ground-truth map of the Yuehai dataset.

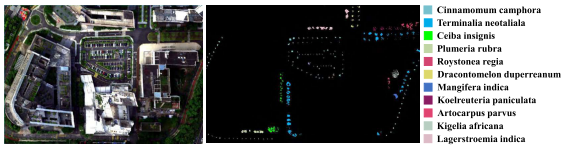

Fig. 3. False-color map and ground-truth map of the Canghai dataset.

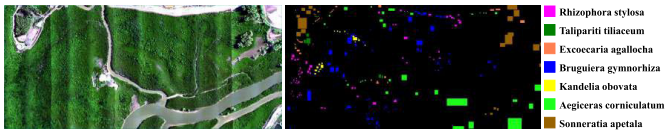

Fig. 4. False-color map and ground-truth map of the Zhanjiang dataset.

Gabor-guided spatial-context self-attention that replaces $\mathbf{K}$ by Gabor to refine image representations better through original prior knowledge. Unlike other models combined with hand-crafted features, the Gabor plays an auxiliary role in our model to guide self-attention to focus on the global spatial context of images and reduce its training load.

Concretely, a 3-D Gaussian function is first used to modulate corresponding filters in various scales and orientations by controlling the central frequency of sinusoidal waves f as well as its angles with the z-axis and xy-plane, φ and θ, in the frequency domain. It is shown as follows:

$$\mathbf{G}_{f,\varphi,\theta}(\alpha,\beta,\lambda) = \frac{1}{(2\pi)^{\frac{2}{3}}\sigma^3} \times \exp\left(2\pi\left(\alpha x + \beta y + \lambda z\right)\right) \times \exp\left(-\frac{\alpha^2+\beta^2+\lambda^2}{2\sigma^2}\right) \tag{10}$$

where (α, β, γ) represents the spatial coordinate of central pixels in the 3-D cube. Moreover,

$$\begin{aligned} x &= f\sin\varphi\cos\theta \\ y &= f\sin\varphi\sin\theta \\ z &= f\cos\varphi\cdot\sigma. \end{aligned} \tag{11}$$

Here, σ is the bandwidth. Subsequently, the modulated Gabor filters one by one slide on each band map and produce corresponding Gabor feature cubes. These cubes will be stacked together along depth and reduced to as same as the original image size via principal component analysis (PCA), forming the final Gabor feature $\mathbf{G}$. As a result, our Gabor-guided spatial-wise self-attention is

$$\mathbf{H}_l = \text{softmax}\left(\frac{\widetilde{\mathbf{Q}}_l\mathbf{G}}{\sqrt{d_k}}\right)\widetilde{\mathbf{V}} \tag{12}$$

where

$$\widetilde{\mathbf{Q}}_l = \widetilde{\mathbf{W}}_l^{\mathbf{Q}}\widetilde{\mathbf{H}}_l, \quad \widetilde{\mathbf{V}}_l = \widetilde{\mathbf{W}}_l^{\mathbf{V}}\widetilde{\mathbf{H}}_l. \tag{13}$$

After (8) and (12), $\mathbf{H}_l$ incorporates spectral features from spectral-wise self-attention and spatial ones from Gabor-guided spatial-context self-attention. Dividing joint spectral–spatial feature extraction into two stages, the model can consider salient spectral maps on the whole image domain and leverage original image prior statistics in each band map to highlight regions of interest. Next, these operations will perform N times. In the end, the class token $\boldsymbol{c}$ serves as our classification feature and is processed by a fully connected layer to get the final result.

IV. Experiments

A. Dataset

1) Yuehai Dataset: The Yuehai dataset was captured on the Yuehai campus of Shenzhen University in 2022 using a UAV equipped with a Specim FX10 hyperspectral camera. The Specim FX10 hyperspectral camera images ground objects in 112 continuous bands within the wavelength range of 400–1000 nm. In this experiment, the area around Times Square was intercepted as the experimental dataset, which covers a scene of 2271 × 3000 pixels, with a spatial resolution of 0.1 m per pixel and a spectral resolution of 5.5 nm. 16 types of tree species were labeled through the field tree species survey. The dataset has a large number of species, and the distribution of tree species is intricate, with a high degree of mixing complexity. Its canopy is dense, so the canopy is seriously shaded. Fig. 2 and Table I provide a detailed display of this dataset.

2) Canghai Dataset: The Canghai dataset was captured on the Canghai campus of Shenzhen University in 2022 using a UAV equipped with a Specim FX10 hyperspectral camera. In this experiment, the area around the building of the School of Computer Science and Software was intercepted as the experimental dataset. The dataset contains 2064 × 3203 pixels with a spatial resolution of 0.1 m per pixel and a spectral resolution of 5.5 nm. According to the field survey results, a total of 11 types of tree species were marked. The tree species in this dataset are scattered and irregularly distributed. The trees are relatively small with sparse leaves. Fig. 3 and Table II are the false color map and sample details of the dataset, respectively.

3) Zhanjiang Dataset: The Zhanjiang dataset is the UAV hyperspectral remote-sensing images of the Gaoqiao mangrove forest area in Zhanjiang. The dataset contains 220 bands with a spatial resolution of 0.3 m, a wavelength range of 400–1000 nm, and a spectral resolution of 2.8 nm. The artificially labeled sample data also mainly comes from the results of field investigations. The concentration of tree species in the Zhanjiang mangrove community is relatively high. And its stand conditions are complex. The tree species in the dataset are interlaced, multiple tree species are intermingled, and the boundaries are not obvious. This experiment intercepted an area with a size of 1639 × 3392 (see Fig. 4). The number of samples in each class is explicitly listed in Table III.

B. Experimental Setup

The main hyperparameters affecting the performance of the model include network depth, learning rate design strategy, convolution kernel settings, and many other factors. In the proposed TASAM, the following factors are mainly considered.

The nested spatial pyramid structure extracts the global overall information and local detailed information of tree species from different receptive fields. Since the street tree with the smallest crown is about seven pixels, the smallest height h and width w of the views in the patch pyramid are set to 7. In addition, it extracts the surrounding adjacent pixel information from different receptive fields to increase the separability of the target and the background. As shown in Fig. 1, the final patch sizes in the patch pyramid are 7, 14, and 28. Then, under each input size, an image pyramid structure with different scales is constructed. The scales of the three image pyramid structures are (7, 5, 3), (14, 7, 5), and (28, 14, 7).

The Gabor function is robust to image brightness changes, contrast changes, and pose changes. One of the most common factors is that the expression of the frequency and direction of the Gabor function is very similar to the human visual response, and Gabor filters with different directions and different scales can extract different features in the image. In the TASAM, the tilt angle θ of the Gabor function is selected as $\theta \in \{(\pi/4), (3\pi/4)\}$, and the phase offset angle φ is selected as $\varphi \in \{0, (\pi/4), (\pi/2), (3\pi/4)\}$. To enhance the multiscale feature representation, the scale factor f is set to $f \in \{0.5, 0.25, 0.125, 0.0625\}$, resulting in four different scales. According to these parameter settings, a group of Gabor filters can be designed to filter the texture image. Each Gabor filter only allows the texture corresponding to its frequency to pass smoothly, while the energy of other textures is suppressed. The texture features analyzed and extracted from each filter output are used in subsequent tasks.

To reduce computational resources and follow the general setting, the hidden unit parameters of all layers in TASAM are set to 64. During training, an Adam optimizer with learning rate 10^{-3} is used to train the model for 1500 epochs. Moreover, to alleviate fluctuation, the learning rate is multiplied by a decay factor of 0.5 at the 800th and 1300th epochs.

In the experiment, 1% of the labeled samples are randomly selected for each class as the training set, and the remaining samples are used as the test set. The number of training samples and test samples in each dataset is shown in Tables I–III. Moreover, each method is repeated ten times to avoid bias caused by random sampling. Finally, the experimental results use overall accuracy (OA), average accuracy (AA), and Kappa coefficient as evaluation indicators to measure the performance of all methods on the three datasets. OA is the ratio of the number of correctly classified category pixels to the total number of categories. AA is the sum of the classification accuracy of each class divided by the total number of classes. The Kappa coefficient is a ratio that represents the proportion of the error reduction produced by classification compared with completely random classification. In addition, the classification result maps corresponding to all methods are masked by the normalized difference vegetation index (NDVI) [56] value to show only the classification effect of vegetation.

C. Ablation Analysis

To verify the effectiveness and rationality of fusing texture features, patch pyramids, and image pyramids, ablation experiments are implemented and analyzed on three datasets (as shown in Table IV). The data in the table reveals the fact that it is difficult to finely classify complex tree species data only with spectral information. As we can see, the experiments using only the spectral self-attention mechanism get the worst results in these three datasets. Especially on the Yuehai dataset, the OA, AA, and Kappa of the spectral self-attention structure are only 69.07%, 48.18%, and 0.64%, respectively.

In addition, the data in the table shows the contribution of the image pyramid and patch pyramid to the framework. On the Canghai dataset, the classification performance of the image pyramid structure is better than that of the patch pyramid structure. This is because an image pyramid can solve the problem of different scales of different objects by extracting global overall information and local detail information from images of different resolutions. Coincidentally, the tree ages in the Canghai dataset are uneven, so the tree crowns are also of different sizes. For example, the crowns of Cinnamomum camphora and Roystonea regia in this dataset are relatively small, while the crowns of Terminalia neotaliala and Artocarpus parvus are larger. However, the tree species in the Yuehai dataset and the Zhanjiang dataset are scattered, densely covered, and have unclear boundaries. Therefore, the patch pyramid structure as shown in the table may be more suitable for these two datasets. The reason is that the patch pyramid structure can extract the spatial difference information of adjacent pixels within multiple receptive fields. However, in the end, as shown in the fourth row of data for each dataset, the nested spatial pyramid structure that combines the image pyramid and patch pyramid can achieve excellent classification results as expected no matter which dataset. A nested spatial pyramid structure can solve the problem of low contrast in the target space. In addition, the highly integrated spatial pyramid structure can achieve the purpose of improving target distinguishability by extracting multiview multiscale features and local detail features. It has high robustness and strong generalization ability. Finally, by comparing the experimental data in the last two rows of each dataset, it can be found that adding the texture features extracted by Gabor-guided spatial-context self-attention can achieve better classification results. Humans distinguish tree species through the shape and texture of leaves, bark, and flowers. Therefore, the spatial texture features of tree species extracted by Gabor-guided spatial-context self-attention are more beneficial to assist the tree species in fine classification.

D. Comparative Experiment

In this section, comparative experiments are carried out on the hyperspectral datasets of the above three tree species. Moreover, several handcrafted-based methods,

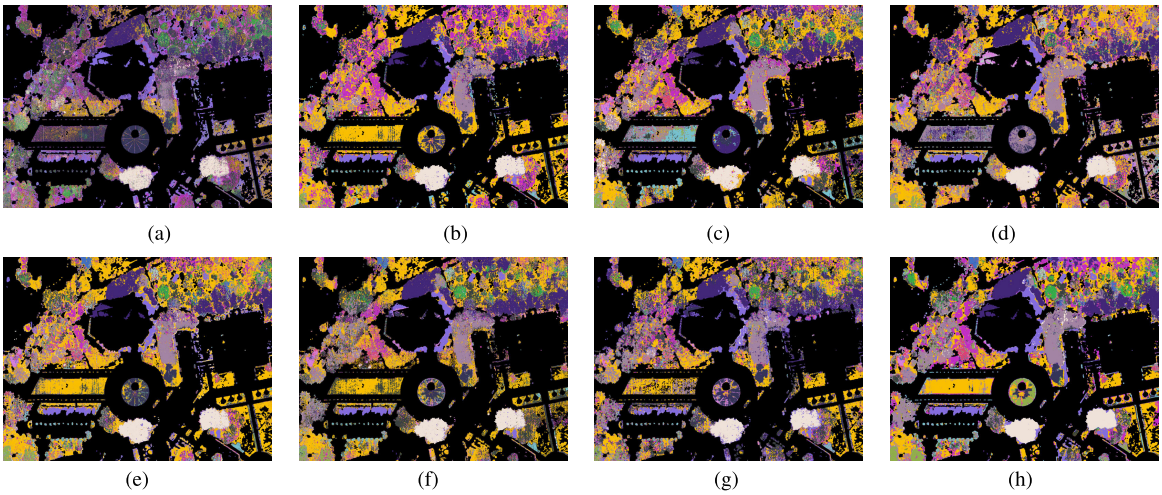

Fig. 5. Classification maps on the Yuehai dataset by (a) SVM (33.72%), (b) CNN (50.52%), (c) 3D-1D-CNN (82.66%), (d) SpectralFormer (57.62%), (e) SCAT (80.27%), (f) SSFTT (71.92%), (g) SSCL3DNN (67.86%), and (h) TASAM (84.49%).

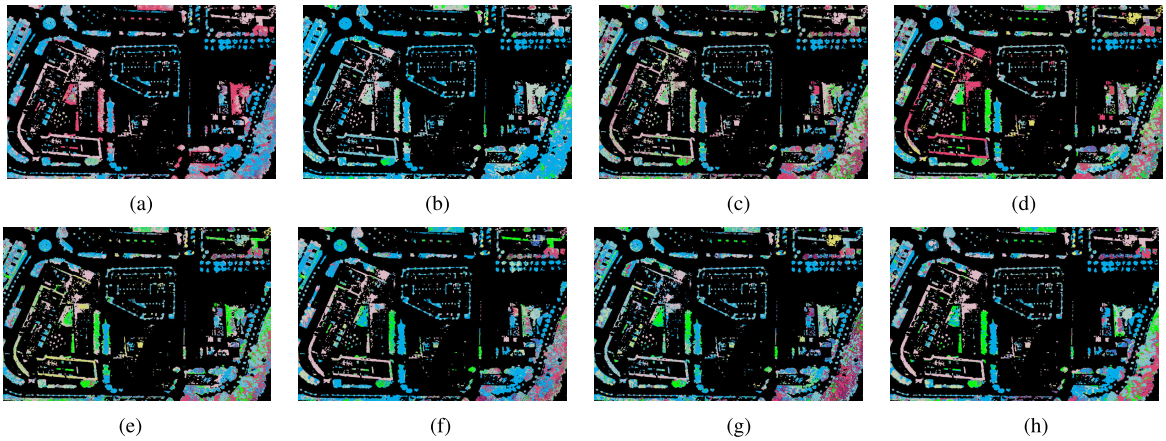

Fig. 6. Classification maps on the Canghai dataset by (a) SVM (46.74%), (b) CNN (63.96%), (c) 3D-1D-CNN (81.84%), (d) SpectralFormer (83.99%), (e) SCAT (88.72%), (f) SSFTT (86.51%), (g) SSCL3DNN (82.95%), and (h) TASAM (90.72%).

TABLE VII

CLASSIFICATION PERFORMANCE BY USING THE SEVEN COMPARED METHODS FOR THE ZHANJIANG HYPERSPECTRAL DATASET WITH 1% LABELED SAMPLES PER CLASS AS THE TRAINING SET (NUMBERS IN BOLD REPRESENT THE BEST CLASSIFICATION PERFORMANCE)

Class	Training	SVM	CNN	3D-1D-CNN	SpectralFormer	SCAT	SSFTT	SSCL3DNN	TASAM
1	249	38.98 ± 7.01	47.59 ± 16.15	60.34 ± 6.46	46.77 ± 6.29	59.27 ± 5.76	71.59 ± 1.51	71.52 ± 1.52	**76.01 ± 2.10**
2	95	59.74 ± 7.50	86.10 ± 3.79	81.77 ± 8.54	86.62 ± 2.21	86.23 ± 2.93	92.88 ± 1.82	83.76 ± 2.70	**92.91 ± 2.39**
3	126	9.66 ± 4.56	13.47 ± 6.67	33.82 ± 18.91	23.64 ± 5.77	29.80 ± 4.78	59.57 ± 6.67	42.94 ± 3.04	**71.54 ± 3.36**
4	580	55.32 ± 16.10	59.44 ± 11.50	74.55 ± 7.81	64.37 ± 6.77	74.65 ± 3.68	86.35 ± 1.36	83.35 ± 1.32	**86.43 ± 2.32**
5	90	12.84 ± 5.40	4.95 ± 6.24	34.84 ± 17.02	19.29 ± 8.20	36.14 ± 12.67	**62.35 ± 3.58**	53.51 ± 3.32	62.01 ± 4.52
6	1013	70.30 ± 10.59	85.20 ± 5.79	87.41 ± 4.23	85.18 ± 3.47	86.80 ± 3.07	**93.08 ± 1.36**	91.02 ± 0.43	91.99 ± 0.74
7	821	83.98 ± 10.93	94.18 ± 4.07	95.06 ± 1.86	92.81 ± 1.73	94.52 ± 0.87	97.23 ± 0.38	94.92 ± 0.46	**98.09 ± 0.32**
OA		47.26 ± 2.53	55.84 ± 1.67	66.83 ± 8.75	59.81 ± 2.87	66.77 ±1.95	80.44 ± 1.13	74.43 ± 0.95	**82.71 ± 1.34**
AA		63.41 ± 4.59	73.51 ± 2.11	80.19 ± 4.56	74.91 ± 2.51	79.74 ± 1.48	88.31 ± 0.42	85.09 ± 0.36	**89.03 ± 0.57**
Kappa		0.51 ± 0.06	0.64 ± 0.03	0.74 ±0.06	0.66 ± 0.03	0.73 ± 0.02	0.85 ± 0.01	0.80 ± 0.00	**0.86 ± 0.01**

including SVM [57], CNN, 3D-1D-CNN [24], SpectralFormer [58], spectral context-aware transformer (SCAT) [59], spectral–spatial feature tokenization transformer (SSFTT) [60], and spatial–spectral ConvLSTM 3-D neural network (SSCL3DNN) [61], are selected as comparison methods. The parameters in all comparison methods are subject to the settings in the original paper.

1) Experimental Results on the Yuehai Dataset: Table V shows the classification results of the proposed TASAM and other comparative methods on the Yuehai dataset. Fig. 5 shows the final classification performance and the corresponding classification maps of different methods on the Yuehai dataset. It can be seen from the table that the OA of the SVM method is only 33.72%, which is 50.77% lower than that of the proposed TASAM method. And except for Araucaria cunninghamii, Ficus elastica, and Terminalia arjuna, which have a large number of training samples, the accuracy of the three classes is more than 80%, and the classification accuracy of other classes is not satisfactory. Perhaps, the primary factor is that SVM only utilizes spectral information and ignores

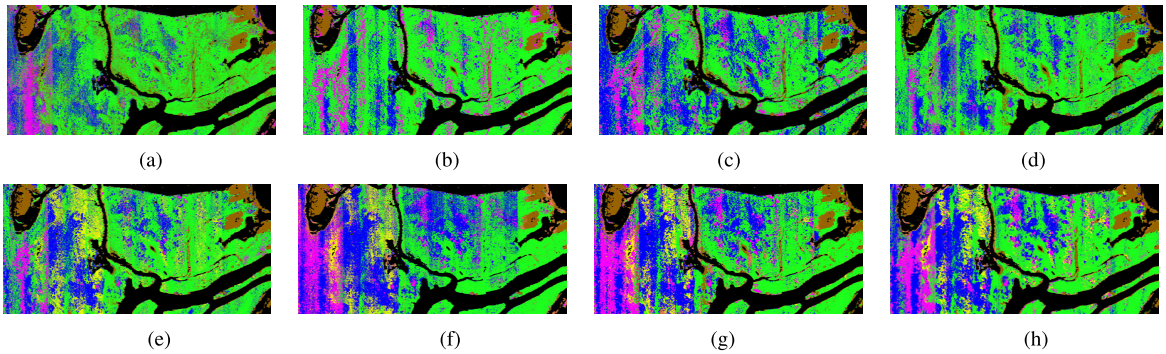

Fig. 7. Classification maps on the Zhanjaing dataset by (a) SVM (47.26%), (b) CNN (55.84%), (c) 3D-1D-CNN (66.83%), (d) SpectralFormer (59.81%), (e) SCAT (66.77%), (f) SSFTT (80.44%), (g) SSCL3DNN (74.43%), and (h) TASAM (82.71%).

spatial information. However, the tree species in the Yuehai dataset are messy, irregularly distributed, and highly mixed, and it is difficult to achieve fine classification only by using spectral information. As illustrated in the classification map, there are many noisy estimates in the SVM classification map. In addition, whether it is OA, AA, or Kappa coefficient, the CNN model and the 3D-1D-CNN model are lower than the proposed TASAM. Although the classification accuracy of the 3D-1D-CNN model is higher than that of the proposed method in a few classes, it does not have much advantage. Moreover, from the classification result map of the CNN method, it is obvious that the Delonix regia represented by green is not recognized. The classification result map of 3D-1D-CNN is relatively better, but the recognition of Pinus is not accurate. There are many reasons responsible for this instance, and the following are the typical ones. Both the CNN model and the 3D-1D-CNN model rely heavily on step-by-step prediction and fail to capture long-distance features.

It is worth noting that the SpectralFormer model also includes an attention module, but its performance on this dataset is far worse than the proposed method. For instance, the OA of SpectralFormer only reaches 57.62%, while the OA of the proposed method is 84.49%. Part of the explanation for it is that SpectralFormer focuses more on the extraction of spectral information, without considering the importance of image spatial geometric information and loses the ability to capture local features. Although the SCAT model considers spatial context information, it has insufficient ability to extract texture features of tree species in images, so the classification effect of SCAT is also inferior to that of the proposed TASAM. For example, the classification accuracies for Roystonea regia and Kigelia africana are only 63.80% and 63.90%, respectively. From the classification results maps of SpectralFormer and SCAT methods, it can also be observed that the location of Roystonea regia and Kigelia africana has not been accurately identified. The performance of the SSFTT method and SSCL3DNN method is also mediocre. Most of the tree species are misidentified in their classification result map. The proposed TASAM can solve the above problems, so no matter what kind of tree species it is, it can achieve better classification accuracy and outperform other comparative models. As we can see, the OA, AA, and Kappa of the proposed TASAM reach 84.49%, 89.83%, and 0.88%, respectively, which are higher than other methods. Moreover, for Leucaena leucocephala with a small number of training samples, the classification accuracy also reached 75.82%. Many factors may account for this result, but the following are the most typical ones. First, the proposed TASAM can not only highlight the discriminative features between different tree species through the nested spatial pyramid module, but also directly calculate the correlation between each pixel through the spectral-context self-attention module. Second, the proposed model also incorporates the Gabor-guided spatial-context self-attention module, which can extract the texture features in the image. The last, but not least, reason is that under the condition of small samples, the proposed model can still fully extract the spatial context information to achieve the ideal effect.

2) Experimental Results on the Canghai Dataset: Table VI and Fig. 6, respectively, show the classification accuracy and classification result maps of the proposed method and other comparison methods on the Canghai dataset. Although the distribution of tree species in this dataset is scattered, there are many types. On this dataset, the SVM method not only has a low OA, but also has a low classification accuracy for each type of tree species. The reason is that SVM mainly supports binary classification and has difficulties in solving multiclassification problems. As illustrated in Fig. 6, in the SVM method classification result map, a large number of tree species are misclassified and identified. The classification performance of the CNN is better than that of SVM, but still worse than that of 3D-1D-CNN. It can be seen that several types of tree species are hardly identified by the CNN method. The OA of 3D-1D-CNN can reach 81.84%, and each class can also achieve a good classification accuracy. However, the overall situation is not better than the proposed method. This conclusion can also be seen from their corresponding classification result maps. The reason for this is that the CNN network and the 3D-1D-CNN network only use convolution operations, which cannot be applied to other images with different spatial structures, so it is not ideal for tree species classification tasks with rich textures.

The classification accuracy of the SpectralFormer method on Terminalia neotaliala reached 96.15%, which is the highest among all methods. However, the classification accuracy of the proposed method on this class also reaches 95.47%, and the classification performance of the proposed method on other classes is significantly better than that of the SpectralFormer

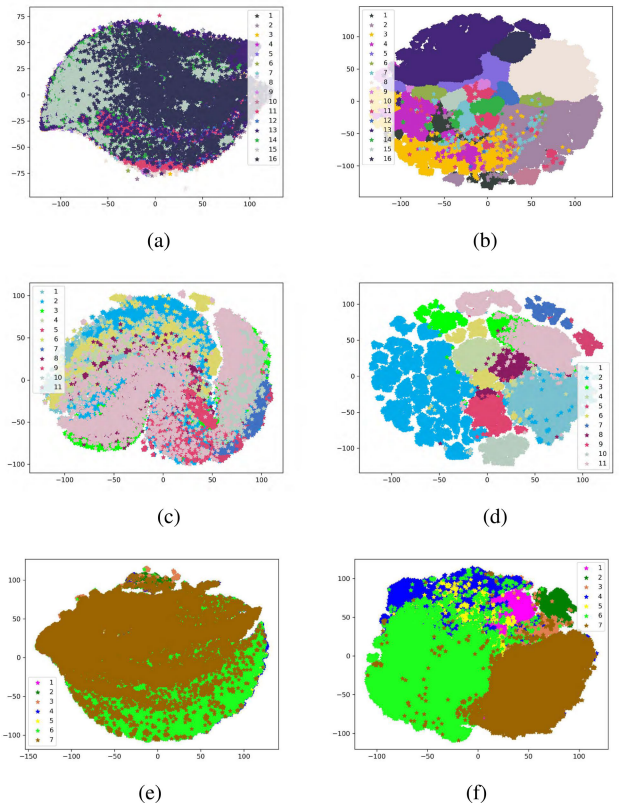

Fig. 8. Sample distributions on three tree species hyperspectral datasets. The first row is the original sample distribution of the three hyperspectral datasets. The second row is the feature distribution of the proposed TASAM on three hyperspectral datasets. (a) Yuehai (original). (b) Yuehai (TASAM). (c) Canghai (original). (d) Canghai (TASAM). (e) Zhanjiang (original). (f) Zhanjiang (TASAM).

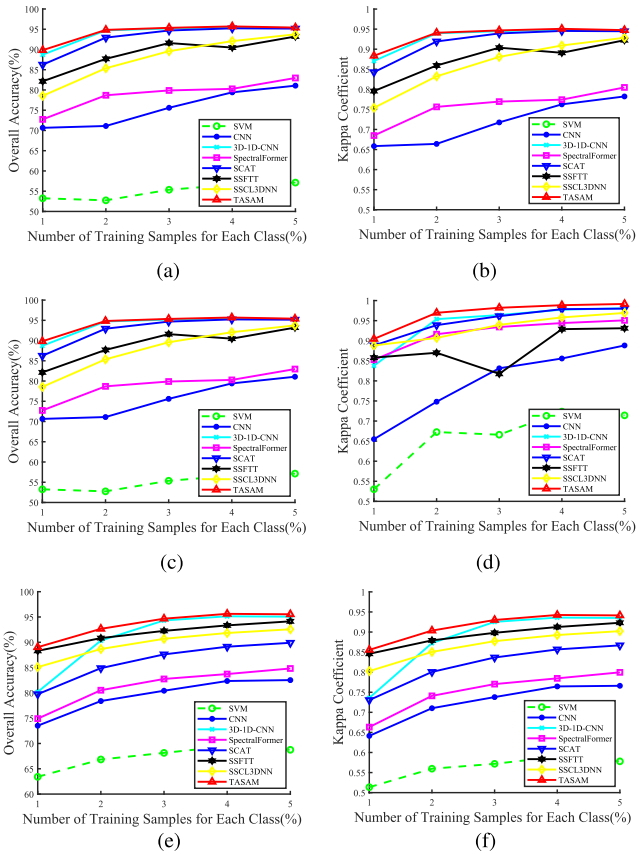

Fig. 9. Classification accuracy for each class with different numbers of training samples on three tree species hyperspectral datasets. (a) Yuehai (OA). (b) Yuehai (Kappa). (c) Canghai(OA). (d) Canghai (Kappa). (e) Zhanjiang (OA). (f) Zhanjiang (Kappa).

method. It can also be found from their corresponding classification result maps that for Dracontomelon duperreanum and Koelreuteria paniculata with similar leaf shapes, the SpectralFormer method does not identify them as accurately as the proposed method. The results indicate that the SpectralFormer framework ignores the spatial structure information of the target, so the classification results for different tree species with different spatial structures are not ideal. The SCAT framework takes each band as an input patch, which can enhance the extraction of spatial context information. Even so, the evaluation indicators of the SCAT framework on different types of tree species are similar to those of the SpectralFormer framework, but also lower than those of the proposed TASAM. The SSFTT method and the SSCL3DNN method also have the same problems mentioned above, so their classification accuracy is not very ideal. In particular, for Roystonea regia, which has the fewest training samples, the classification accuracy of the SSCL3DNN method is only 63.09%. The OA, AA, and Kappa of the proposed TASAM in this dataset are all superior to other comparison methods and reach 90.72%, 92.15%, and 0.90, respectively. It is worth pointing out that the proposed TASAM can achieve high classification accuracy for tree species with complex distribution and crown structure when the number of training samples is small. There are many reasons for this phenomenon. The main reason is that the proposed TASAM can utilize a 3-D Gabor filter to extract the texture features of multiple types of tree species from different directions and different scales so that it has higher adaptability to tree species with complex texture structures. Overall, the table shows that the proposed TASAM has better generalization and robustness for fine classification of numerous tree species with scattered and irregular distributions.

3) Experimental Results on the Zhanjiang Dataset: The results of comparative experiments on the Zhanjiang dataset are shown in Table VII and Fig. 7. As shown in the table, the classification performance of the SVM method is still slightly inferior. The OA of the SVM method is only 47.26%, and the classification accuracy for Excoecaria agallocha is only 9.66%. The classification result map of the SVM method also shows that many tree species are misidentified. Furthermore, it can be noticed that the classification performance of the CNN method and the 3D-1D-CNN method on this dataset is very general. The explanation for this is that the pooling layer of the CNN method and the 3D-1D-CNN method will lose certain valuable information, ignoring the correlation between the local and the global. The OA of the SpectralFormer method and SCAT method also reached only 59.81% and 66.77%, respectively. It is fairly easy to find out the reason that the SCAT framework can extract spectral information as well as spatial context information, but still cannot achieve better results. The reason is that different tree species have huge differences in canopy structure and leaf shape, and their spatial

texture structure also varies greatly, but the SCAT framework ignores this information. For example, the leaves of Aegiceras corniculatum grow in pairs, and the flowers also grow in pairs. Its leaves are obovate, with flat midribs and slightly raised lateral veins. But the leaves of Kandelia obovata are nonoval, with inconspicuous veins and thick petioles. It can also be seen from the results that the recognition accuracy of the SpectralFormer method and SCAT method for Kandelia obovata is only 19.29% and 36.14%, respectively, and Kandelia obovata is hardly seen in the classification result map [see Fig. 7(d)] of SpectralFormer method.

The only surprising thing is that the SSFTT method has the highest classification accuracy on the two classes Kandelia obovata and Aegiceras corniculatum. Because the SSFTT method combines the convolution module with the Transformer structure to extract shallow spatial–spectral features in HSIs. However, the SSFTT method can only have satisfactory performance on this dataset. For the above two tree species datasets with intricate distribution and unclear boundaries, the classification performance is not ideal. The proposed method can also achieve satisfactory results on this dataset, especially since the classification accuracy of Sonneratia apetala has reached 98.09%. Even though the dataset contains shrubs and trees, and the species are mixed. The nested spatial pyramid module of the proposed method can still extract multilevel feature maps from multiscale images of different receptive fields, thereby improving the contrast and identification of tree species. Moreover, a cross-spectral–spatial attention module is also introduced to extract spectral and texture features to further enhance the differences between tree species. Therefore, the proposed TASAM can not only achieve fine classification on this dataset, but also acquire perfect classification accuracy on the above two datasets.

E. Analysis of Sample Distributions

Fig. 8 shows the sample distribution map of the original tree species HSI and the sample distribution map after feature extraction by the proposed TASAM. It can be seen from Fig. 8(a), (c), and (e) that the original sample distributions in the three tree species HSIs have a high degree of overlap after being projected into the 2-D space. Especially in the images of the Yuehai and Zhanjiang hyperspectral datasets, only the sample distribution of two or three classes can be vaguely seen. Comparing the two images of Fig. 8(a) and (b), it can be found that the features of the original image are not discriminative, so the sample distributions of different classes have a large overlap. After the feature extraction of the proposed method, the sample features have better discrimination, so samples of different classes are separated from each other, and samples of the same class are aggregated. The most obvious is the separation of the third class and the sixth class. In addition, Fig. 8(f) shows that regardless of the number of samples in each class, they can be separated by the proposed method. The distribution of tree species in the Canghai tree species hyperspectral dataset is relatively scattered. Therefore, although the sample distribution in the original image is not as high as the other two datasets, it is also chaotic. After the multiview and multiscale feature extraction and cross-spectral–spatial attention feature extraction are performed on the data by the proposed method, the samples clearly show strong separability in the feature space. As shown in Fig. 8(d), all classes are almost perfectly separated. In conclusion, the visualization results further demonstrate that the proposed TASAM can effectively improve the separability between samples.

F. Classification Results With Different Numbers of Training Samples

In this section, we analyze the effect of using different numbers of training and testing samples on the classification results of different methods. Fig. 9 shows the OA and Kappa coefficients of the proposed TASAM method and other comparative methods for the three tree species HSI datasets under various sample conditions. The number of training samples will directly affect the classification accuracy of the classifier. As the number of training samples increases, the classification accuracy will gradually rise. However, the proposed TASAM method can always obtain the highest classification accuracy of tree species, especially under the condition of small samples, it always outperforms other comparative methods. For example, when the number of training samples is 2% on the Zhanjiang dataset, the lowest classification accuracy among the comparison methods is the SVM method, which can only reach 66.84%. In addition, the two comparison methods with the highest classification accuracy are the SSFTT method and the 3D-1D-CNN method, and their classification accuracy is 90.83% and 90.33% respectively. Nevertheless, the proposed TASAM achieves the best classification accuracy at this time and is 92.66%. Moreover, regardless of the number of training samples, the Kappa coefficient of the proposed method is the highest, which further proves the effectiveness and stability of the proposed method.

V. Conclusion

This article proposes a texture-aware self-attention mechanism for hyperspectral tree species classification. In the proposed TASAM network, a nested spatial pyramid module that can accurately extract multiview and multiscale features concerning objects is constructed to the discriminative features of tree species and improve the contrast between tree species and the surrounding background at an early stage. In addition, a cross spectral–spatial attention comprised of spectral-context self-attention and Gabor-guided spatial-context self-attention is designed to capture joint spectral–spatial features so that our model can improve spatial contrast and overcome spectral variance further. Concretely, the spectral-context self-attention extracts spectral correlation on the whole image domain, dynamically adapting the variance happening in different stages. Furthermore, Gabor features serve as auxiliary roles in the spatial context, guiding the attention mechanism to focus on latent spatial texture features and further enhancing discriminability. In summary, our model can effectively combine spectral features and texture features to improve the classification accuracy of hyperspectral tree species. Meanwhile, it can also enhance the distinction between the phenomena of "different objects with the same spectrum" and "same objects

with different spectra." Ablation experiments and comparative experiments on three tree species hyperspectral datasets demonstrate the effectiveness, robustness, and generalization of our model with limited labeled samples. This method can greatly reduce the workload of forestry personnel and meet the application requirements of forestry resource investigation and forestry carbon sink analysis.

REFERENCES

[1] K. He, "China's carbon neutrality faces the challenges of 'three highs and one short', and requires 'five carbon implementations' to achieve dual carbon goals," *iEnergy*, vol. 2, no. 1, pp. 2–3, Mar. 2023.

[2] Y. Su, "Understandings of carbon peaking, carbon neutrality, and energy development strategy of China," *iEnergy*, vol. 1, no. 2, pp. 145–148, Jun. 2022.

[3] J. Guo, S. Ma, T. Wang, Y. Jing, W. Hou, and H. Xu, "Challenges of developing a power system with a high renewable energy proportion under China's carbon targets," *iEnergy*, vol. 1, no. 1, pp. 12–18, Mar. 2022.

[4] G. Steur, R. W. Verburg, M. J. Wassen, and P. A. Verweij, "Shedding light on relationships between plant diversity and tropical forest ecosystem services across spatial scales and plot sizes," *Ecosyst. Services*, vol. 43, Jun. 2020, Art. no. 101107.

[5] L. Huang, M. Zhou, J. Lv, and K. Chen, "Trends in global research in forest carbon sequestration: A bibliometric analysis," *J. Cleaner Prod.*, vol. 252, Apr. 2020, Art. no. 119908.

[6] H. Zhou, C. Yan, and H. Huang, "Tree species identification based on convolutional neural networks," in *Proc. 8th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, vol. 2, Aug. 2016, pp. 103–106.

[7] D. E. Guyer, G. E. Miles, L. D. Gaultney, and M. M. Schreiber, "Application of machine vision to shape analysis in leaf and plant identification," *Trans. ASAE*, vol. 36, no. 1, pp. 163–171, 1993.

[8] M. Molinier and H. Astola, "Feature selection for tree species identification in very high resolution satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 4461–4464.

[9] 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.

[10] S. L. Sutton, "Alice grows up: Canopy science in transition from wonderland to reality," *Plant Ecol.*, vol. 153, nos. 1–2, pp. 13–21, 2001.

[11] M. A. Hasan et al., "Temporal changes in land cover, Land Surface Temperature, soil moisture, and evapotranspiration using remote sensing techniques—A case study of Kutupalong Rohingya refugee camp in Bangladesh," *J. Geovisualization Spatial Anal.*, vol. 7, no. 1, p. 11, Jun. 2023.

[12] J. Lee et al., "Individual tree species classification from airborne multisensor imagery using robust PCA," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2554–2567, Jun. 2016.

[13] F. E. Fassnacht et al., "Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2547–2561, Jun. 2014.

[14] X. Liao, B. Tu, J. Li, and A. Plaza, "Class-wise graph embedding-based active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522813.

[15] L. Chen, Y. Wei, Z. Yao, E. Chen, and X. Zhang, "Data augmentation in prototypical networks for forest tree species classification using airborne hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410116.

[16] R. Richter, B. Reu, C. Wirth, D. Doktor, and M. Vohland, "The use of airborne hyperspectral data for tree species classification in a species-rich central European forest area," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 52, pp. 464–474, Oct. 2016.

[17] P. Gong, "Conifer species recognition: An exploratory analysis of in situ hyperspectral data," *Remote Sens. Environ.*, vol. 62, no. 2, pp. 189–200, Nov. 1997.

[18] M. E. Martin, S. D. Newman, J. D. Aber, and R. G. Congalton, "Determining forest species composition using high spectral resolution remote sensing data," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 249–254, Sep. 1998.

[19] W. Koedsin and C. Vaiphasa, "Discrimination of tropical mangroves at the species level with EO-1 hyperion data," *Remote Sens.*, vol. 5, no. 7, pp. 3562–3582, Jul. 2013.

[20] D. Harrison, B. Rivard, and A. Sánchez-Azofeifa, "Classification of tree species based on longwave hyperspectral data from leaves, a case study for a tropical dry forest," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 66, pp. 93–105, Apr. 2018.

[21] T. Hycza, K. Stereńczak, and R. Bałazy, "Potential use of hyperspectral data to classify forest tree species," *New Zealand J. Forestry Sci.*, vol. 48, no. 1, pp. 1–13, Dec. 2018.

[22] G. A. Fricker, J. D. Ventura, J. A. Wolf, M. P. North, F. W. Davis, and J. Franklin, "A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery," *Remote Sens.*, vol. 11, no. 19, p. 2326, Oct. 2019.

[23] J. Cao, W. Leng, K. Liu, L. Liu, Z. He, and Y. Zhu, "Object-based mangrove species classification using unmanned aerial vehicle hyperspectral images and digital surface models," *Remote Sens.*, vol. 10, no. 2, p. 89, Jan. 2018.

[24] B. Zhang, L. Zhao, and X. Zhang, "Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111938.

[25] F. Tong and Y. Zhang, "Spectral–spatial and cascaded multilayer random forests for tree species classification in airborne hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411711.

[26] Z. Guo, M. Zhang, W. Jia, J. Zhang, and W. Li, "Dual-concentrated network with morphological features for tree species classification using hyperspectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7013–7024, Aug. 2022.

[27] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[28] L. Liu, Y. Pang, W. Fan, Z. Li, and M. Li, "Fusion of airborne hyperspectral and LiDAR data for tree species classification in the temperate forest of northeast China," in *Proc. 19th Int. Conf. Geoinformatics*, Jun. 2011, pp. 1–5.

[29] D. Zhao, Y. Pang, L. Liu, and Z. Li, "Individual tree classification using airborne LiDAR and hyperspectral data in a natural mixed forest of northeast China," *Forests*, vol. 11, no. 3, p. 303, Mar. 2020.

[30] J. Mäyrä et al., "Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks," *Remote Sens. Environ.*, vol. 256, Apr. 2021, Art. no. 112322.

[31] D. S. W. Katz, S. A. Batterman, and S. J. Brines, "Improved classification of urban trees using a widespread multi-temporal aerial image dataset," *Remote Sens.*, vol. 12, no. 15, p. 2475, Aug. 2020.

[32] R. Ahmed, K. H. Mahmud, and J. H. Tuya, "A GIS-based mathematical approach for generating 3D terrain model from high-resolution UAV imageries," *J. Geovisualization Spatial Anal.*, vol. 5, no. 2, pp. 1–10, Dec. 2021.

[33] 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.

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

[35] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2536–2551, 2023.

[36] Y. Chu, J. Fei, and S. Hou, "Adaptive global sliding-mode control for dynamic systems using double hidden layer recurrent neural network structure," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1297–1309, Apr. 2020.

[37] M. Fetanat, M. Stevens, P. Jain, C. Hayward, E. Meijering, and N. H. Lovell, "Fully Elman neural network: A novel deep recurrent neural network optimized by an improved Harris Hawks algorithm for classification of pulmonary arterial wedge pressure," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1733–1744, May 2022.

[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] R. Xin, J. Zhang, and Y. Shao, "Complex network classification with convolutional neural network," *Tsinghua Sci. Technol.*, vol. 25, no. 4, pp. 447–457, Aug. 2020.

[40] J. Huang, S. Huang, Y. Zeng, H. Chen, S. Chang, and Y. Zhang, "Hierarchical digital modulation classification using cascaded convolutional neural network," *J. Commun. Inf. Netw.*, vol. 6, no. 1, pp. 72–81, Mar. 2021.

[41] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[43] Y. Qing, W. Liu, L. Feng, and W. Gao, "Improved transformer net for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 11, p. 2216, Jun. 2021.
[44] B. Tu, X. Liao, Q. Li, Y. Peng, and A. Plaza, "Local semantic feature aggregation-based transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536115.
[45] Y. Gao, M. Zhou, and D. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 61–71.
[46] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 14–24.
[47] L. He, C. Liu, J. Li, Y. Li, S. Li, and Z. Yu, "Hyperspectral image spectral–spatial-range Gabor filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4818–4836, Jul. 2020.
[48] B. Zhang, Y. Aziz, Z. Wang, L. Zhuang, M. K. Ng, and L. Gao, "Hyperspectral image stripe detection and correction using Gabor filters and subspace representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[49] M. Idrissa and M. Acheroy, "Texture classification using Gabor filters," *Pattern Recognit. Lett.*, vol. 23, no. 9, pp. 1095–1102, Jul. 2002.
[50] J. Kim, S. Um, and D. Min, "Fast 2D complex Gabor filter with kernel decomposition," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1713–1722, Apr. 2018.
[51] J. Y. Choi and B. Lee, "Ensemble of deep convolutional neural networks with Gabor face representations for face recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3270–3281, 2020.
[52] C. Liu, J. Li, L. He, A. Plaza, S. Li, and B. Li, "Naive Gabor networks for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 376–390, Mar. 2021.
[53] S. Jia et al., "3-D Gabor convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509216.
[54] S. Jia, X. Deng, J. Zhu, M. Xu, J. Zhou, and X. Jia, "Collaborative representation-based multiscale superpixel fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7770–7784, Oct. 2019.
[55] S. Jia et al., "Flexible Gabor-based superpixel-level unsupervised LDA for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10394–10409, Dec. 2021.
[56] N. Ghasemloo, A. A. Matkan, A. Alimohammadi, H. Aghighi, and B. Mirbagheri, "Estimating the agricultural farm soil moisture using spectral indices of Landsat 8, and Sentinel-1, and artificial neural networks," *J. Geovisualization Spatial Anal.*, vol. 6, no. 2, pp. 1–12, Dec. 2022.
[57] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[58] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
[59] N. Li, J. Xue, and S. Jia, "Spectral context-aware transformer for cholangiocarcinoma hyperspectral image segmentation," in *Proc. 5th Int. Conf. Image Graph. Process. (ICIGP)*, 2022, pp. 209–213.
[60] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.
[61] W.-S. Hu, H.-C. Li, L. Pan, W. Li, R. Tao, and Q. Du, "Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4237–4250, Jun. 2020.

Nanying Li received the B.E. degree in automation and the M.E. degree in information and communication engineering from the Hunan Institute of Science and Technology, Yueyang, China, in 2017 and 2021, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include hyperspectral image classification, anomaly detection, and image segmentation.

Shuguo Jiang received the B.E. degree from Xiamen University of Technology, Xiamen, China, in 2020, and the M.E. degree from Shenzhen University, Shenzhen, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China.

His research interests include remote sensing and multimodality interpretation.

Jiaqi Xue received the B.E. degree in software engineering from the Taiyuan University of Technology, Shanxi, 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 classification, deep learning, and multioperator fusion.

Songxin Ye received the B.E. 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.

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.