Progressive Semantic Enhancement Network for Hyperspectral and LiDAR Classification

Xiyou Fu, *Member, IEEE*, Xi Zhou, Yawen Fu, Pan Liu, and Sen Jia, *Senior Member, IEEE*

***Abstract*— The joint classification of hyperspectral image (HSI) and light detection and ranging (LiDAR) data is gaining attention for its improved classification accuracy. However, effectively integrating the rich spectral information of HSI and the elevation features of LiDAR has remained a challenge in multimodal fusion. This article proposes a novel approach called progressive semantic enhancement network (PSENet) for hyperspectral and LiDAR classification based on a progressive joint spatial–spectral attention mechanism. PSENet mainly comprises two modules: the spatial grouping constraint (SAGC) module and the spectral weighting constraint (SEWC) module. The SAGC module extracts multiscale features in the spatial domain, while the SEWC module focuses on enhancing semantic features in spectral dimension. By gradually utilizing spatial and spectral constraint modules to progressively enhance feature extraction, PSENet integrates affluent information for a more refined classification of ground objects. Based on experimental results, it has been demonstrated that PSENet outperforms several most advanced methods on three datasets. The SAGC and SEWC modules proposed in PSENet enable the effective integration of the spatial, spectral, and elevation information from HSI and LiDAR, providing a promising way to perform classification more accurately. The source codes of this work will be publicly available at http://szu-hsilab.com/.**

***Index Terms*— Attention mechanism, fusion classification, hyperspectral image (HSI), light detection and ranging (LiDAR).**

I. Introduction

IN RECENT years, the classification of ground objects has become increasingly important in remote sensing (RS). Among various RS data, hyperspectral image (HSI) stands out due to its rich spectral information, which can significantly aid in object classification [1], [2]. Hyperspectral imaging is a primary technique in Earth observation and has found extensive application across various fields, such as urban planning [3], [4], environmental monitoring [5], [6], precision agriculture [7], [8], change detection [9], [10], and anomaly detection [11], [12]. However, despite the rapid development of the classification techniques for HSI, we still cannot achieve the desired classification accuracy in some applications, due to the susceptibility of HSI to atmospheric changes [13], and the phenomenon that the same object has different spectra and different objects have identical spectra [14].

Light detection and ranging (LiDAR) technology differs from hyperspectral imaging in that it uses laser light for active ranging. This technique is less affected by weather conditions and can provide accurate elevation information, enabling comprehensive acquisition of the spatial characteristics of ground objects [15]. However, LiDAR technology primarily relies on single-wavelength laser detection and cannot capture the spectral information of objects. This limitation results in a limited ability to classify ground objects in complex scenes using LiDAR data alone [16]. Therefore, data from a single sensor typically have limitations in certain applications [17]. Currently, a lot of research focuses on the fine recognition of ground objects through the fusion of complementary information from multiple RS data sources. Due to the different working principles of Earth observation sensors, data obtained by various equipment can reflect different characteristic details of the ground objects. For example, by fusing hyperspectral and LiDAR data, it is possible to leverage the complementary advantages of spatial, spectral, and elevation information, thus enhancing the precision of identifying objects [18].

The traditional classification approaches of multimodal RS data involve extracting morphological features followed by selected classifiers such as support vector machine (SVM) [19] and random forest (RF) [20] for fusion. Khodadadzadeh et al. [21] proposed a multifeature learning strategy for fusing and classifying hyperspectral imagery and LiDAR data, extracting attribute profiles from both sources. The advantage of not requiring any regularization parameters makes it an effective way to utilize and integrate different types of features. Ghamisi et al. [22] proposed the extinction profile (EP) method to extract spatial features of HSI more efficiently, resulting in higher classification accuracy than AP based on the extinction filter. Rasti et al. [23] proposed an orthogonal total variational analysis method to fuse the features of HSI and LiDAR followed by the classification based on the EP method for feature extraction. However, traditional approaches have some shortcomings when dealing with HSI and LiDAR data. First, these methods require extensive expertise for manual feature design in the RS domain. Since HSI and LiDAR

Received 27 November 2023; revised 31 July 2024 and 29 September 2024; accepted 29 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42301375 and Grant 62271327; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515110076 and Grant 2022A1515011290; in part by Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, Grant JCYJ20240813141635047, and Grant KQTD20200909113951005; and in part by the Research Team Cultivation Program of Shenzhen University under Grant 2023JCT002. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: fuxy0623@szu.edu.cn; 992566968@qq.com; 2200271038@email.szu.edu.cn; lp2548836497@163.com; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3513979

data have different feature expressions, feature extraction is more challenging. Second, these methods usually use simple feature fusion methods such as splicing and summation, which cannot fully utilize the information advantages of HSI and LiDAR data for better fusion. Third, most of these methods use shallow models such as linear classifiers or decision trees, making it difficult to improve classification accuracy.

With the emergence of deep learning (DL), particularly the rapid advancements in convolutional neural networks (CNNs), more performance-efficient DL methods have emerged for the joint classification using HSI and LiDAR data [24], [25]. DL methods can automatically learn features, extract rich feature information from multimodal RS data, and are suitable for dealing with complex nonlinear relationships that may exist in RS data. Therefore, DL methods have become one of the mainstream methods for joint classification using HSI and LiDAR data. Chen et al. [26] designed two separate CNNs to extract the respective features of HSI and LiDAR data and then used a fully connected (FC) neural network to fuse the features obtained by the two CNNs for classification. Inspired by the end-to-end idea, Chen et al. [27] directly fused HSI and LiDAR data at the pixel level of the original data at first. Then, they used two CNNs to extract spectral and spatial features that were finally superimposed and classified. However, this approach may result in insufficient information mining of LiDAR data. Zhang et al. [28] proposed a patch-to-patch CNN to combine multiscale features between different modalities for collaborative classification of the two-modal data. Roy et al. [29] extended the traditional self-attention mechanism by introducing cross-modal self-attention modules to classify hyperspectral and LiDAR data. To address the issue of pixelwise features that are ineffective with current cascade or weighted fusion approaches, Lu et al. [30] first trained a coupled adversarial feature learning network to learn higher order semantic features in an unsupervised manner and then performed classification through a supervised multilevel feature fusion. Sun et al. [31] designed an end-to-end lightweight network based on CNN and Transformer for hyperspectral and LiDAR data classification. However, these DL-based methods may not be able to capture the change of logical representation relations hidden in the semantic information of two modalities from shallow to deep during the fusion process. They often ignore the weight constraints of different semantic information at different depths of the network, leading to the loss of relevant spectral information from HSI and elevation information from LiDAR during the fusion process.

The use of attention mechanisms has become increasingly crucial in image classification [32]. Essentially, attention mechanisms allow DL models to better focus on essential features or areas, reducing the interference of irrelevant information [33]. However, most works only focus on processing single-modal data in multiscale or multimodal data in single-scale, without taking into consideration that multimodal data may bear different contributions at various scales. Based on this, we propose a progressive semantic enhancement network (PSENet) with the objective of gradually enhancing the semantic information from shallow features to deep features during the feature fusion process. The aim is to accurately analyze the spatial distribution and trends of the Earth's surface, thereby uncovering the intrinsic information and characteristics of RS data and reducing classification errors.

To summarize, the main contributions of PSENet are given as follows.

1) In the PSENet, we propose a spatial grouping constraint (SAGC) module that incorporates a multiscale cross-modal spatial attention mechanism to enhance the spatial properties of land objects. This mechanism strengthens the multiscale spatial information by incorporating elevation from LiDAR data at different group levels, thus improving the distinguishability of land objects that exhibit high similarity at a single scale.
2) We propose a spectral weighting constraint (SEWC) module to capture the intrinsic properties of land objects. It adaptively enhances spectral weights based on interdependence among channels in multiscale spatial features. Group convolutions emphasize spectral interaction within the same scale, followed by feature fusion across scales, allowing for better integration of spatial, spectral, and elevation information.
3) Unlike other parallel joint spatial–spectral attention mechanisms commonly used in the classification, the two modules proposed in PSENet progressively constrain and enhance the semantic information as the network deepens. By utilizing spatial and spectral constraint modules to progressively enhance features, PSENet integrates affluent information for a more refined classification of ground objects.

We organize this article as follows. Section II provides a brief overview of multimodal RS data fusion and attention mechanisms in DL. Section III presents the proposed model and algorithm in detail. Section IV describes the experimental setup and analysis of the results. Finally, Section V presents the conclusion of this article.

II. Related Works

A. Fusion of HSI and LiDAR Data

The purpose of fusing HSI and LiDAR data is to extract the most useful feature information from complementary and redundant data sources to form a more comprehensive description of the scene. This fusion enables more accurate classification because of the comprehensive description of the scene. From a methodological perspective, the fusion of HSI and LiDAR data can be broadly categorized into pixel-, feature-, and decision-level fusions [34].

Generally, pixel-level fusion refers to directly averaging HSI and LiDAR data at the pixel level [35]. While pixel-level fusion preserves more information, it results in a high computational burden. Feature-level fusion maximizes the extraction of meaningful features from HSI and LiDAR data, offering higher computational efficiency and compressing feature information with minimal loss. Compared to pixel-level fusion, feature-level fusion involves fewer parameters, resulting in improved processing efficiency. Decision-level fusion involves analyzing the overall results using logical operations and decision rules. However, it requires more data preprocessing and feature extraction techniques and heavily relies on

classification outcomes [36]. It is worth noting that due to significant differences in imaging modalities between HSI and LiDAR data, fusion at the feature level is the most widely used method [37].

The morphological feature concatenation fusion methods have been widely used in the early classification of HSI and LiDAR data. However, this approach often leads to high-dimensional features, significantly increasing the computational complexity of subsequent classification steps [38]. To address the issue of redundant information in morphological features, fusion methods based on low-rank models have emerged [23]. These methods aim to transform features from a high-dimensional space to a low-dimensional space. Another fusion method is based on composite kernels, which utilizes kernel space mapping for feature dimensionality reduction and fusion [39]. To leverage global similarity measures among samples, graph-based fusion methods utilize graph space for feature representation. Among these approaches, the work described in [40] is a typical example, which establishes corresponding topological graphs for different features to achieve graph space mapping. However, a drawback of such methods is that computing the similarity between samples becomes time-consuming when dealing with large-scale images. With the impressive capabilities demonstrated by DL, methods based on DL have also gained significant attention in HSI and LiDAR data fusion. Representative models in this context include single-level feature fusion structures and multilevel feature fusion structures based on CNNs [41].

B. Attention Mechanism

In DL, the attention mechanism is an important technique that enables the neural network to focus more on specific input data [42]. Typically, neural networks consider all information equally when processing input data, but, in some cases, certain parts of the information may be more important to the successful completion of the task at hand. The attention mechanism calculates the importance weight of each input data and then applies these weights to the calculation process of the neural network, allowing the network to focus more on important elements in the input sequence. Researchers have developed many plug-and-play modules based on the attention mechanism, which have improved the performance of related tasks in the natural image field. For instance, Hu et al. [43] proposed the squeeze-and-excitation attention mechanism to selectively modulate the scale of channels for capturing channel correlations. However, the proposed attention mechanism does not consider attention in the spatial dimension. To address this shortcoming, Woo et al. [44] combined convolution and attention mechanisms to pay attention to images from both spatial and channel aspects. Subsequently, Wang et al. [45] proposed an efficient mechanism based on channel attention, in which a local cross-channel interaction strategy and a method to adaptively determine the coverage of local cross-channel interaction are introduced. Recently, due to its ability to effectively integrate and associate information from different modalities, enhancing overall understanding and performance, cross-modal attention has been proposed and applied in many fields, such as image segmentation [46], data fusion [47], and object detection [48].

The advantages of the attention mechanism for image processing have been verified in the field of natural images [49], [50]. Inspired by this, researchers have introduced it into the field of RS, hoping to enhance the model's ability to process spatial, spectral, and elevation information from HSI and LiDAR data [51], [52]. Mohla et al. [53] proposed a feature extraction and fusion framework for land cover classification from HSI and LiDAR data. The proposed framework uses the self-attention mechanism to highlight the spectral features of HSI and emphasizes the spatial features of HSI using the cross-attention mechanism by employing the attention map of LiDAR data. Fang et al. [54] used the cross-attention mechanism instead of the self-attention mechanism for cross-modal information interaction. Similarly, cross-modal attention modules have also been proposed in [29] and [31] to classify hyperspectral and LiDAR data. These works demonstrate the importance and effectiveness of the attention mechanism in HSI and LiDAR classification and provide useful inspiration for research in related fields. However, existing cross-modal attention mechanisms in hyperspectral and LiDAR classification are typically single-scale, ignoring the spatial difference of ground objects at various levels, which is precisely the intrinsic characteristics of ground objects and is crucial for classification.

III. Methodology

In order to fully utilize the complementary advantages of HSI and LiDAR data, we intend to strengthen the distinguishing ability of the classification model from the perspectives of spatial and spectral domains. Therefore, the proposed model presented in this section mainly consists of two components: 1) an SAGC module and 2) an SEWC module. In terms of the spatial dimension, SAGC can comprehensively describe the distribution characteristics of various ground objects under different scale structures. Regarding the spectral dimension, SEWC can capture and emphasize the category information of ground features. In the following, we will illustrate the detailed architecture of the proposed model.

A. Network Structure

Given the HSI $\boldsymbol{X}_h \in \mathbb{R}^{M\times N\times C_h}$ and LiDAR $\boldsymbol{X}_l \in \mathbb{R}^{M\times N\times C_l}$ obtained at the same area of the Earth, where M and N represent the height and width of the two images, and C_h and C_l denote the number of channels in the HSI and LiDAR, respectively, consider each pixel $Y = \{y^i\}_{i=1}^n$ of the image as the center, and select HSI and LiDAR patches $X = \{\boldsymbol{x}_h^i, \boldsymbol{x}_l^i\}_{i=1}^n$ into the proposed model. Here, $\boldsymbol{x}_h \in \mathbb{R}^{P\times P\times C_h}$ and $\boldsymbol{x}_l \in \mathbb{R}^{P\times P\times C_l}$, where P and n represent the size of the neighborhood and the number of pixel samples available in the image, respectively. Labels $y_i^n \in \{1, 2, \ldots, K\}$, where K represents the number of label categories. Next, these patches are fed into the model for training. The architecture of PSENet is shown in Fig. 1. Due to the mismatch in channel dimensions between the original HSI and LiDAR data, we have devised two separate feature extraction modules to extract convolutional features of

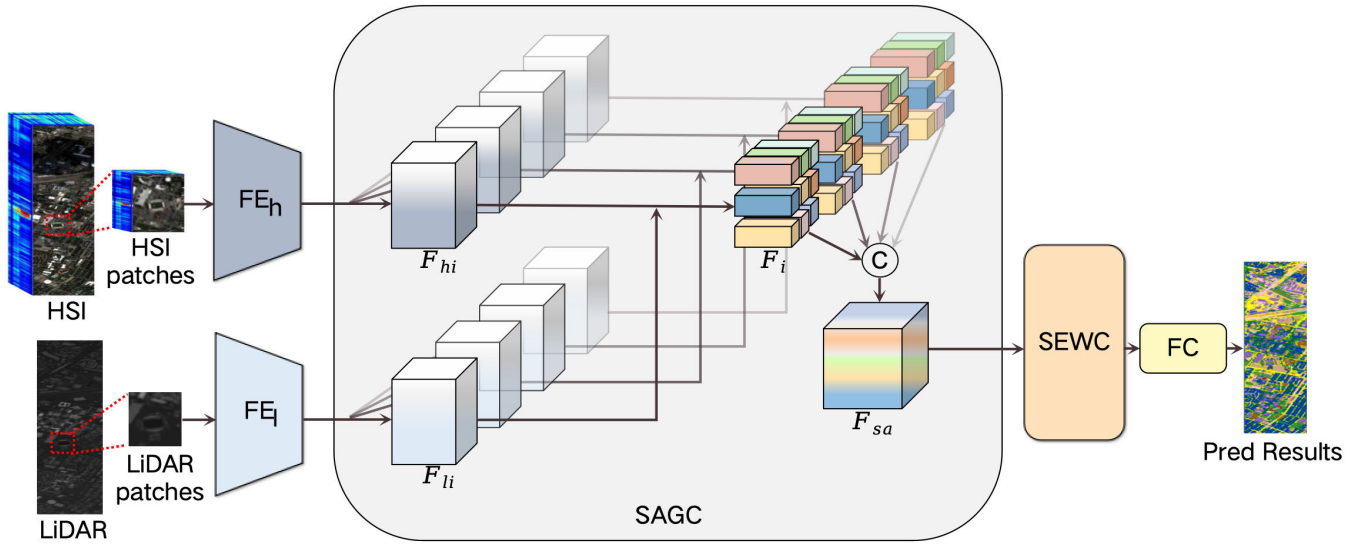

Fig. 1. Illustration of the proposed PSENet framework. The network is mainly composed of two modules, SAGC and SEWC, which are proposed for adaptive semantic information constraints in spatial and spectral dimensions, respectively.

the same dimensions for these two modalities. These feature extraction modules are denoted as $FE_h(\cdot)$ and $FE_l(\cdot)$, and they yield the extracted features $\boldsymbol{F}_h$ and $\boldsymbol{F}_l$, respectively. Then, the feature extraction operation can be formulated as

$$\boldsymbol{F}_h = FE_h(\boldsymbol{x}_h) \tag{1}$$

$$\boldsymbol{F}_l = FE_l(\boldsymbol{x}_l). \tag{2}$$

Let Conv$(\cdot)$, $BN(\cdot)$, and ReLU$(\cdot)$ denote convolutional layer, batch normalization, and ReLU activation functions, respectively. The two feature extraction modules share the same structure and can also be more specifically expressed as

$$FE_h(\cdot) = \text{ReLU}(BN(\text{Conv}(\text{ReLU}(BN(\text{Conv}(\cdot)))))) \tag{3}$$

$$FE_l(\cdot) = \text{ReLU}(BN(\text{Conv}(\text{ReLU}(BN(\text{Conv}(\cdot))))))). \tag{4}$$

After extracting certain abstract features, the SAGC module and the SEWC module based on the joint attention strategy are used to adaptively group multiscale spatial information and enhance the spectral information. Let SAGF$(\cdot)$ and SEWF$(\cdot)$ represent SAGC and SEWC, respectively. The formula is expressed as

$$\boldsymbol{F}_{sa} = \text{SAGC}(\boldsymbol{F}_h, \boldsymbol{F}_l) \tag{5}$$

$$\boldsymbol{F}_f = \text{SEWC}(\boldsymbol{F}_{sa}). \tag{6}$$

Finally, the fused features are input into the classification module for obtaining pixel-by-pixel classification results.

B. SAGC Module

Due to the limitation of single-scale features in effectively expressing interclass differences and distinguishing object boundaries in the classification, the SAGC module is proposed to address this issue by grouping the spatial information of the input intermediate feature maps using a multibranch structure. This ensures that the fused features carry rich information from various spatial scales. In addition, this cross-modal spatial attention allows for exploring the effective complementary elevation information provided by LiDAR to enhance the spatial and spectral features of HSI at different scales, thereby compensating for the limited information conveyed by HSI data. Considering the need to learn features at different spatial scales, we set the groups of branches $T = 4$ in the SAGC module, each capturing features at a different scale. Then, the multiscale spatial feature maps can be obtained by channel concatenation, denoted as Cat$(\cdot)$. This operation combines the feature maps from different branches along the channel dimension, resulting in a comprehensive representation that encompasses information from multiple scales, which can be formulated as

$$\boldsymbol{F}_{sa} = \text{Cat}([\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4]) \tag{7}$$

where $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$, and $\boldsymbol{F}_4$ represent the feature maps outputted by the four groups of branches, each capturing features at a different scale. $\boldsymbol{F}_{sa}$ refers to the multiscale spatial feature map generated by the SAGC module, which is obtained by combining and integrating the feature maps from these different branches.

The branch structure of SAGC is shown in Fig. 2. Specifically, in the channel dimension, we generate four groups of HSI and LiDAR features with different receptive fields using convolutional filters of varying sizes, i.e., 3×3, 5×5, 7×7, and 9×9. Let us assume that the dimensions of the intermediate feature maps for both HSI and LiDAR data are C, and each feature map at a different scale has a common channel dimension of $C' = (C/T)$. This ensures that each scale captures distinct spatial information while maintaining consistency in the channel dimension. The HSI and LiDAR features at the i scales, i.e., $\boldsymbol{F}_{hi}$ and $\boldsymbol{F}_{li}$, are expressed as

$$\boldsymbol{F}_{hi} = \text{ReLU}(BN(\text{Conv}(\boldsymbol{F}_h))), i = 1, 2, 3, 4 \tag{8}$$

$$\boldsymbol{F}_{li} = \text{ReLU}(BN(\text{Conv}(\boldsymbol{F}_l))), i = 1, 2, 3, 4. \tag{9}$$

At this time, each location of HSI and LiDAR features in space contains information of the same range, and the proposed

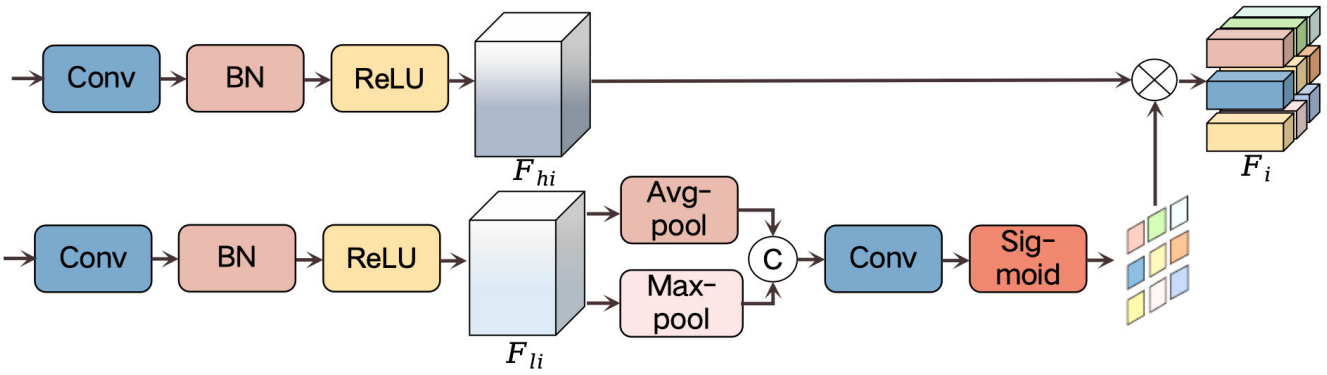

Fig. 2. Illustration of one branch of the SAGC module. SAGC is a module containing four branches, achieving spatial grouped attention across modalities at different scales.

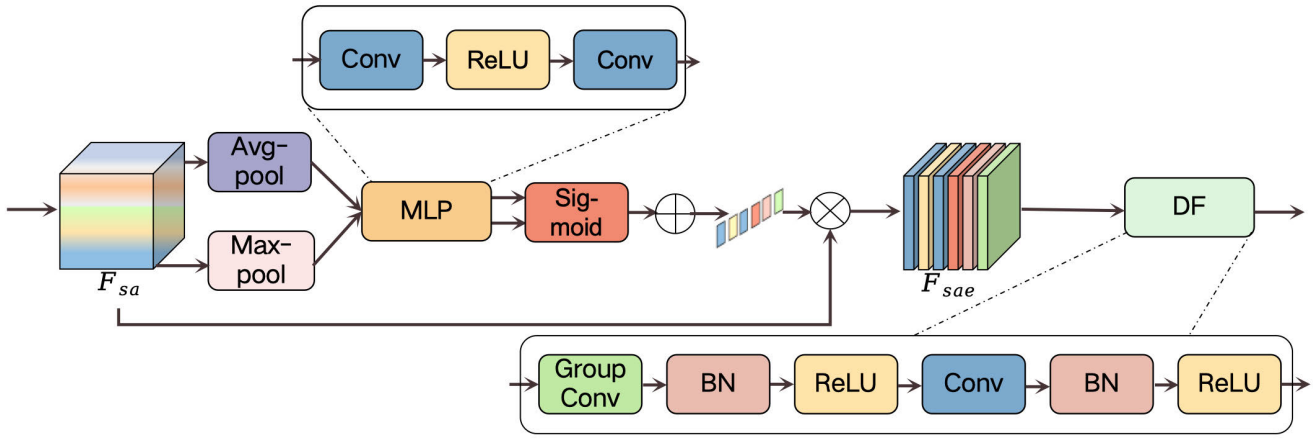

Fig. 3. Illustration of SEWC. SEWC fuses features through spectral attention, grouped convolution, and ordinary convolution.

model enhances the representation ability of different ground objects. Then, LiDAR spatial attention maps of corresponding scales are generated by different branches. Let AvgPool(·) and MaxPool(·) mean the operation of average pooling and maximum pooling, respectively, Sigmoid(·) represents the sigmoid activation function, and the formula is expressed as

$$\boldsymbol{F}_{mi} = \text{Cat}(\text{AvgPool}(\boldsymbol{F}_{li}), \text{MaxPool}(\boldsymbol{F}_{li})), i = 1, 2, 3, 4 \tag{10}$$

$$\boldsymbol{F}'_{li} = \text{Sigmoid}(\text{Conv}(\boldsymbol{F}_{mi})), i = 1, 2, 3, 4. \tag{11}$$

Among them, $\boldsymbol{F}_{mi}$ is the corresponding feature map of each branch after the average, maximum pooling, and concatenation operation, and $\boldsymbol{F}'_{li}$ is the spatial attention map finally obtained by different branches. The spatial attention map is created by using average pooling and maximum pooling operations across different scales of feature maps. This helps to aggregate channel information and generate corresponding weight coefficients. By multiplying the spatial attention maps with the features of HSI at different scales, the constraints of height information can be adaptively added to the features of HSI. The formula is expressed as

$$\boldsymbol{F}_i = \boldsymbol{F}'_{li} \otimes \boldsymbol{F}_{hi}, i = 1, 2, 3, 4. \tag{12}$$

Under the constraint of the SAGC, the proposed model places greater emphasis on the pixel regions that shadow significant impact on classification, effectively disregarding less relevant areas. However, relying solely on the spatial extent of land cover attributes is insufficient. As the model learns deeper features, the model tends to extract more high-level semantic information. Therefore, it is necessary to highlight the essential attributes of land cover from the spectral dimension, which are more relevant for classification tasks.

C. SEWC Module

The previous module focused only on the spatial multiscale constraints between the two modalities. With the deepening of the network, the semantic information between different channels of HSI becomes more distinct. At this point, the deeper semantic information becomes more abstract and no longer requires spatial information constraints. Therefore, SEWC utilizes attention strategies to train genuine spectral weights on multiscale features, enhancing the feature representation ability of key feature channels. This enables PSENet to progressively and adaptively process deeper semantic information and enhance the spectral representation of features, thereby improving the ability to accurately distinguish ground objects.

The structure of SEWC is shown in Fig. 3. To eliminate spatial disturbances, we first employ average pooling and max pooling operations on the multiscale feature maps to obtain channel-level global features. Performing these operations in

parallel ensures a more comprehensive information extraction. Next, the resulting feature maps from both pooling operations are individually fed into a multilayer perceptron (MLP), and their outputs are passed through activation functions to obtain weighted feature vectors. Finally, the weighted features from both branches are elementwise summed to infer finer channel attention. The formula is expressed as

$$\mathrm{MLP}(\cdot) = \mathrm{Conv}(\mathrm{ReLu}(\mathrm{Conv}(\cdot))) \tag{13}$$

$$\boldsymbol{F}_{\mathrm{avg}} = \mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F}_{sa}))) \tag{14}$$

$$\boldsymbol{F}_{\mathrm{max}} = \mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}_{sa}))) \tag{15}$$

$$\boldsymbol{F}_{se} = \boldsymbol{F}_{\mathrm{avg}} + \boldsymbol{F}_{\mathrm{max}} \tag{16}$$

where $\boldsymbol{F}_{\mathrm{avg}}$ and $\boldsymbol{F}_{\mathrm{max}}$ represent the weighted feature vectors obtained by average pooling and max pooling operations, respectively. $\boldsymbol{F}_{se}$ denotes the final feature vector obtained by summing these two vectors, which represents the recalibrated channel weights in the spectral attention map. MLP$(\cdot)$ refers to the MLP layer. Note that in order to reduce model complexity and improve generalization, the MLP layers for $\boldsymbol{F}_{\mathrm{avg}}$ and $\boldsymbol{F}_{\mathrm{max}}$ can be considered the same, composed of two shared-parameter convolutional layers. This design allows for shared learning across the two branches.

Third, the recalibrated HSI channel weights are elementwise multiplied with the multiscale feature map, yielding a refined multiscale feature map with enriched spectral information. The formula is expressed as

$$\boldsymbol{F}_{sae} = \boldsymbol{F}_{sa} \otimes \boldsymbol{F}_{se} \tag{17}$$

where $\boldsymbol{F}_{sae}$ represents the output multiscale refined feature map. At this stage, the proposed model exhibits enhanced discriminative power across different channels, achieving the objective of focusing on specific feature map channels that are crucial for classification.

Finally, under spectral constraints, we perform deeper fusion using convolution, which can be divided into grouped convolution and regular convolution. The purpose of grouped convolution is to reduce the volume of the parameter while isolating the information exchange between different groups in the refined feature maps. In other words, the recalibrated feature maps initially emphasize the intrinsic information of different-scale feature maps. Then, through regular convolution, the spectral and spatial information exchange between different groups is achieved, resulting in the final deep fusion. The formula is expressed as

$$DF(\cdot) = \mathrm{ReLU}(BN(\mathrm{Conv}(\mathrm{ReLU}(BN(\mathrm{GConv}(\cdot)))))) \tag{18}$$

$$\boldsymbol{F}_f = DF(\boldsymbol{F}_{sae}) \tag{19}$$

where $\boldsymbol{F}_f$ represents the final feature output by the SEWC module. $DF(\cdot)$ represents the deep fusion operation, and GCONV$(\cdot)$ represents the grouped convolution. After the cross-channel information interaction, the feature $\boldsymbol{F}_f$ is fed into an FC layer for classification, resulting in the final classification prediction map. The FC layer utilizes the learned features to make predictions and assign class labels to the input data based on the extracted information.

TABLE I
NUMBER OF TRAINING AND TESTING SAMPLES FOR EACH LAND COVER CLASS ON THE SZUTREE DATASET

Class No.	Land Cover Type	No. of Samples	Training	Testing
1	Ficus	653	10	643
2	Ficus microcarpa	6935	10	6925
3	Litchi	13861	10	13851
4	Hoop pine	10978	10	10968
5	Acacia auriculaeformis	34032	10	34022
6	Camphor tree	4296	10	4286
	Total	70755	60	70695

TABLE II
NUMBER OF TRAINING AND TESTING SAMPLES FOR EACH LAND COVER CLASS ON THE HOUSTON2013 DATASET

Class No.	Land Cover Type	No. of Samples	Training	Testing
1	Health grass	1251	10	1241
2	Stressed grass	1254	10	1244
3	Synthetic grass	697	10	687
4	Trees	1244	10	1234
5	Soil	1242	10	1232
6	Water	325	10	315
7	Residential	1268	10	1258
8	Commercial	1244	10	1234
9	Road	1252	10	1242
10	Highway	1227	10	1217
11	Railway	1235	10	1225
12	Parking lot 1	1233	10	1223
13	Parking lot 2	469	10	459
14	Tennis court	428	10	418
15	Running track	660	10	650
	Total	15029	150	14879

IV. EXPERIMENTS

A. Dataset Description

To validate the effectiveness of the proposed method, we conducted comparative and ablation experiments on three HSI-LiDAR paired datasets: SZUTree, Houston2013, and MUUFL Gulfport. The SZUTree dataset, containing six tree species, was constructed to evaluate the accuracy of tree species classification.

1) SZUTree Dataset: The dataset was captured at the Canghai Campus of Shenzhen University, Shenzhen, China, by an unmanned aerial vehicle (UAV). The HSI data consist of 112 bands with wavelengths ranging from 400 to 1000 μm. The HSI and LiDAR data have a spatial resolution of 10 cm, with a size of 1005 $\times$ 900. The ground-truth samples are distributed into six distinct classes. Table I shows the number of training and testing samples for each land cover class.

2) Houston2013 Dataset: This dataset was acquired in and around the University of Houston campus and was featured in the 2013 GRSS Data Fusion Competition [40]. The HSI data consist of 144 bands ranging from 0.38 to 1.05 μm in the wavelength. Both the HSI and LiDAR data have a spatial resolution of 2.5 m and dimensions of 349 $\times$ 1905. The ground-truth samples are classified into 15 distinct categories. Table II shows the number of training and testing samples for each land cover class.

3) MUUFL Gulfport Dataset: The dataset was captured in November 2010 [55], which initially contained 72 bands; 64 bands were used after excluding the first and last four due to noise. The LiDAR data consist of two elevation gratings.

TABLE III
NUMBER OF TRAINING AND TESTING SAMPLES FOR EACH LAND COVER CLASS ON THE MUUFL GULFPORT DATASET

Class No.	Land Cover Type	No. of Samples	Training	Testing
1	Trees	23246	10	23236
2	Grass pure	4270	10	4260
3	Grass groundsurface	6882	10	6872
4	Dirt and sand	1826	10	1816
5	Road Materials	6687	10	6677
6	Water	466	10	456
7	Building's shadow	2233	10	2223
8	Buildings	6240	10	6230
9	Sidewalk	1385	10	1375
10	Yellow curb	183	10	173
11	Cloth panels	269	10	259
	Total	53687	110	53577

All bands and rasters were registered to obtain a total size of 325 × 220. There are a total of 53 687 ground-truth pixels, including 11 categories. Table III shows the number of training and testing samples for each land cover class.

B. Compared Methods

Nine methods were used as comparison methods in this study, including CNN-HSI [56], CoupledCNNs [57], S^{2}ENet [54], FusAtNet [53], Early-Fusion [41], Middle-Fusion [41], LateFusion [41], Cross-HL [29], and CALC [30]. CNN-HSI is a representative CNN-based single-source classification method. CoupledCNNs is a classic framework for fusing HSI and LiDAR data by using two coupled CNNs. The fusion methods of Early-Fusion, Middle-Fusion, and Late-Fusion models involve simple feature concatenation at the shallow, middle, and deep layers of the CNN, respectively. S^{2}ENet, FusAtNet, Cross-HL, and CALC are state-of-the-art models that use attention mechanisms for HSI and LiDAR classification. We found that most of the models were trained using different platforms. To ensure the comparability of experimental results, we implemented all the experiments in a rigorous and consistent experimental environment. All methods were trained and tested on the PyTorch platform.

C. Experimental Settings and Evaluation Metrics

All the experiments mentioned in this article were carried out using an Intel Xeon Silver 4314 CPU and NVIDIA GA102GL [A40] graphics card using the PyTorch framework. To overcome the challenge of acquiring large labeled training data in reality, only ten labeled samples were randomly selected for each class in the three datasets presented in this article for model training. During the training process, the proposed model was optimized using the Adam algorithm, with the cross-entropy function serving as the loss function. We assessed classification performance using three metrics: overall accuracy (OA), average accuracy (AA), and kappa coefficient. To avoid the influence of random factors and better demonstrate the model's stability, we averaged the experimental results. Specifically, all experiments, including the comparative experiments and ablation experiments, were conducted using the same training and test samples based on the random division of the dataset, with the final performance obtained by the average of the results in ten repeating experiments.

D. Experimental Results and Analysis

1) SZUTree Dataset: Table IV presents the experimental results of different methods on the SZUTree dataset, with the best results highlighted in bold for clarity. The single-source classification model CNN-HSI has a notably lower classification accuracy on the tree species dataset, with an OA 26.63% lower than that of the proposed model. All competing classification models based on the fusion of two RS data, i.e., HSI and LiDAR, have achieved good fusion performance on the SZUTree dataset. Among them, CoupleCNNs, Cross-HL, and CALC obtain comparable classification performance. S^{2}ENet achieves the best classification performance except for the proposed method. Among the three different fusion stages, Middle-Fusion achieved good fusion performance, with a 2.90% higher OA than Early-Fusion and a 1.70% higher OA than Late-Fusion. It can be inferred that among the three simple fusion methods, the SZUTree dataset is not suitable for the early fusion stage and the late fusion stage. Early-Fusion, being an early fusion method, may result in a mismatch of features between HSI and LiDAR data, potentially forcing the network to fuse certain information. On the other hand, Late-Fusion, as a late fusion method, may lead to excessive abstraction of information and the potential loss of a significant amount of detailed information.

Fig. 4 displays the classification maps of the competitors, the ablation experiments, and the proposed method for the SZUTree dataset. In the classification result map of CNN-HSI, there are many misclassified areas for the ficus (class 1) and the acacia auriculaeformis (class 5), while the results of other multisource classification models that have included LiDAR data appear closer to the ground truth. It is evident that compared with other methods, the tree species categories predicted by the proposed method are in better agreement with the ground truth. This is because the proposed method can effectively integrate all cross-modal information. In other words, subtle texture distinctions of different tree species at various scales, as well as the elevation changes, can be effectively captured by multiscale spatial attention maps obtained from LiDAR data.

2) Houston2013 Dataset: Table V compares the classification accuracies of the proposed method and competitors for the Houston2013 dataset. Surprisingly, the OA of the FusAtNet model is only 76.93%, which is 4.30% lower than the OA of the single-source classification model CNN-HSI. The OA of the Cross-HL model is only 81.08%. We infer that one of the factors affecting the classification accuracy of the FusAtNet and Cross-HL model on the Houston2013 dataset is the limited training samples. Since the Houston dataset contains more categories and each category occupies fewer pixels, FusAtNet, which directly employs cross-attention, cannot able to learn richer intraclass and interclass relationships when there are fewer labeled samples. Although S^{2}ENet and the proposed method are also based on the attention mechanism, they both introduce a feature extraction module to enhance the information representation of HSI and LiDAR before using the attention mechanism for feature fusion. CALC achieved the second-best classification result with an OA of 89.31%, which is a bit lower than 90.09% of the proposed method.

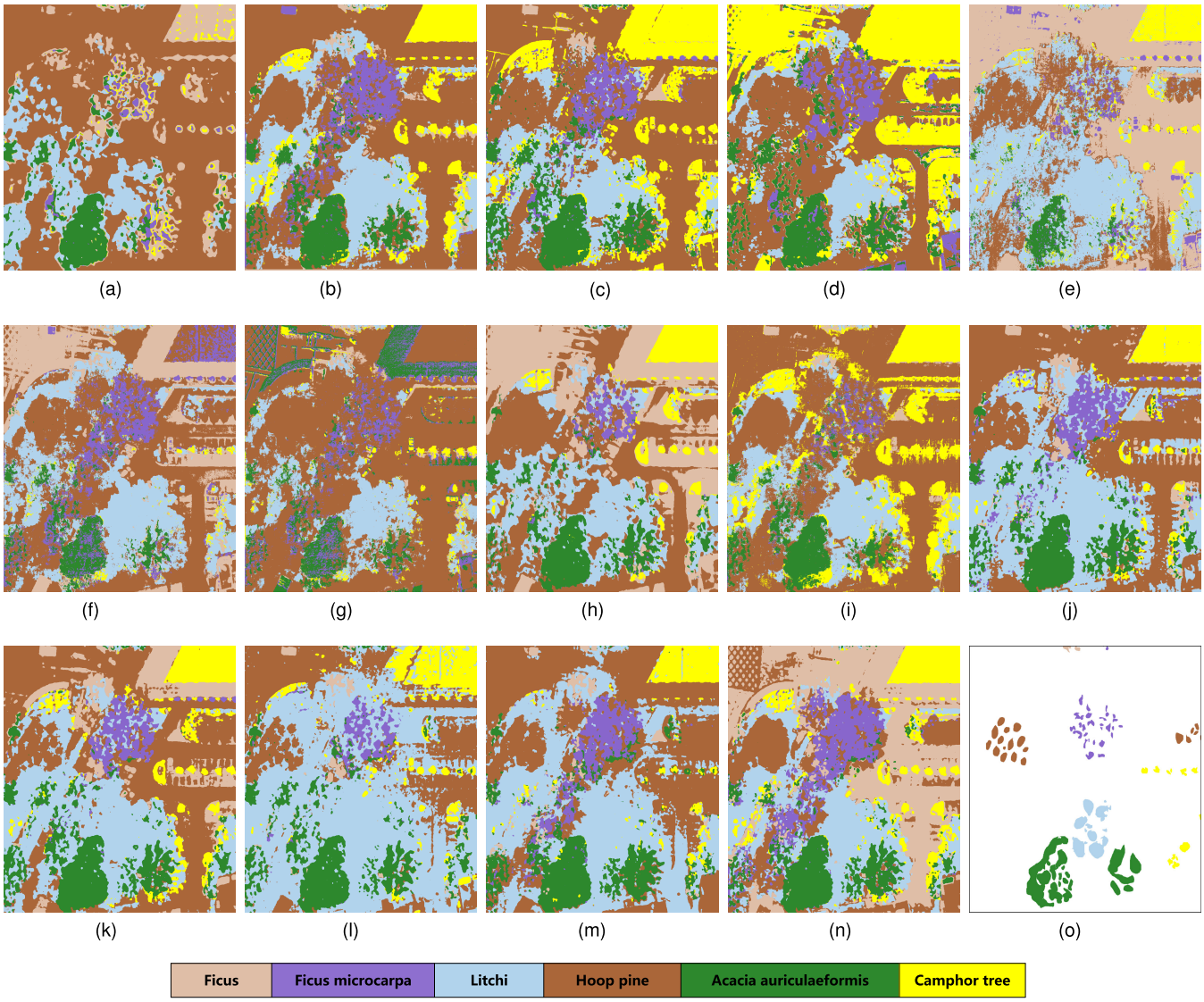

Fig. 4. Classification maps and ground-truth map for the SZUTtee dataset. (a) CNN-his (70.28%). (b) CoupledCNNs (90.98%). (c) S^{2}ENet (93.20%). (d) FusAtNet (84.07%). (e) Early-fusion (84.08%). (f) Middle-fusion (86.98%). (g) Late-fusion (85.28%). (h) Cross-HL (89.94%). (i) CALC (90.44%). (j) w/o SEWC (93.16%). (k) w/o SAGC (93.57%). (l) SAGC$_{_1G}$ (92.97%). (m) SAGC$_{_2G}$ (93.50%). (n) PSENet (96.91%). (o) Ground-truth map.

TABLE IV
CLASSIFICATION ACCURACY FOR THE SZUTREE DATASET

Class No.	CNN-HSI	CoupledCNNs	S^{2}ENet	FusAtNet	Early-Fusion	Middle-Fusion	Late-Fusion	Cross-HL	CALC	PSENet
1	40.54	93.55	71.96	71.31	53.56	77.01	72.08	66.16	**98.76**	86.10
2	59.89	89.56	92.82	70.26	67.06	80.23	68.12	85.11	68.12	**97.88**
3	89.36	96.85	94.99	92.59	96.23	96.75	96.28	95.04	**99.91**	98.14
4	59.16	99.32	97.46	87.67	97.32	91.80	90.56	99.45	99.81	**99.87**
5	73.40	86.68	92.45	83.86	80.08	83.39	83.57	89.65	84.47	**96.39**
6	33.53	86.73	86.27	73.19	74.81	83.90	79.57	62.70	**96.59**	89.47
OA(%)	70.28	90.98	93.20	84.07	84.08	86.98	85.28	89.94	90.44	**96.91**
AA(%)	59.31	92.11	89.38	79.81	78.18	85.51	81.70	83.02	93.50	**94.64**
Kappa	0.5959	0.8736	0.9029	0.7757	0.7777	0.8180	0.7926	85.57	86.70	**0.9555**
Time(s)	2298.48	**328.11**	2030.76	12346.95	1994.94	2197.92	2568.45	3318.90	4182	3149.10

Fig. 5 displays the classification maps of the competitors, the ablation experiments, and the proposed method for the Houston2013 dataset. It is evident that the CNN-HSI model misclassifies the highway (class 10) as other classes. Similar to the situation in the SZUTree dataset, this may also be due to the fact that it is a single-source classification model without the assistance of elevation information, making it prone to producing abnormal classification results.

3) MUUFL Gulfport Dataset: Table VI presents the classification accuracies of the proposed method and competitors for the MUUFL Gulfport dataset. S^{2}ENet is also one of the models that achieves the best classification performance, except for

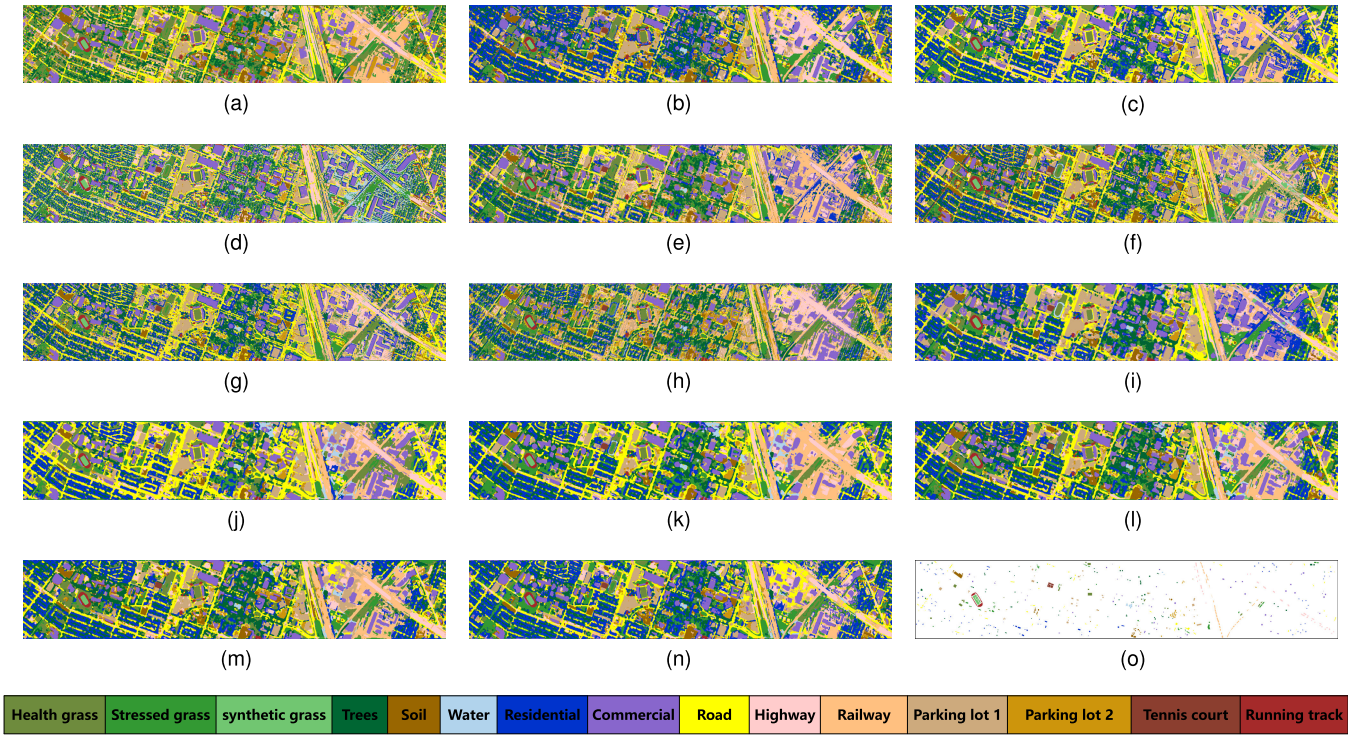

Fig. 5. Classification maps and ground-truth map for the Houston2013 dataset. (a) CNN-his (81.23%). (b) CoupledCNNs (88.35%). (c) S^2ENet (86.30%). (d) FusAtNet (76.93%). (e) Early-fusion (85.42%). (f) Middle-fusion (85.76%). (g) Late-fusion (84.57%). (h) Cross-HL (81.08%). (i) CALC (89.31%). (j) w/o SEWC (89.22%). (k) w/o SAGC (88.65%). (l) $SAGC_{\_1G}$ (88.72%). (m) $SAGC_{\_2G}$ (89.45%). (n) PSENet (90.09%). (o) Ground-truth map.

TABLE V
CLASSIFICATION ACCURACY FOR THE HOUSTON2013 DATASET

Class No.	CNN-HSI	CoupledCNNs	S^2ENet	FusAtNet	Early-Fusion	Middle-Fusion	Late-Fusion	Cross-HL	CALC	PSENet
1	91.56	87.70	91.28	81.50	91.31	87.89	89.39	**92.26**	82.95	90.05
2	92.53	92.11	91.09	81.25	93.14	90.68	90.01	90.42	87.54	**92.70**
3	88.69	96.89	99.55	95.88	97.58	**99.80**	99.04	98.28	96.83	99.36
4	94.10	**97.74**	95.92	92.02	94.64	96.56	94.95	93.27	94.97	95.49
5	99.13	99.20	99.05	89.04	99.40	99.19	97.30	95.20	99.19	**99.74**
6	89.51	91.14	89.90	75.59	85.40	86.67	84.41	85.02	**94.48**	91.97
7	69.12	86.88	86.19	72.67	84.72	85.27	80.98	72.59	90.81	**92.72**
8	55.54	73.87	75.62	65.43	70.40	73.63	74.16	62.79	**80.86**	76.99
9	72.75	75.47	73.72	61.52	73.78	70.30	70.10	69.79	80.92	**82.00**
10	80.52	82.70	70.17	59.33	71.03	70.85	71.69	74.87	**83.21**	79.70
11	71.13	90.54	85.11	72.11	74.66	82.06	81.65	71.76	**95.87**	91.92
12	64.06	79.34	72.02	68.14	79.99	77.60	72.04	61.95	79.89	**84.12**
13	80.45	95.40	93.86	79.46	92.48	93.88	93.33	84.62	**95.99**	93.33
14	97.41	99.67	99.33	96.94	98.30	99.19	97.56	95.31	**99.95**	99.86
15	**99.89**	99.75	99.77	92.18	98.42	99.88	98.86	97.78	99.35	99.72
OA(%)	81.23	88.35	86.30	76.93	85.42	85.76	84.57	81.08	89.31	**90.09**
AA(%)	83.09	89.89	88.16	78.87	87.02	87.58	86.36	83.06	90.85	**91.31**
Kappa	0.7973	0.8741	0.8519	0.7508	0.8423	0.8461	0.8332	0.7955	0.8845	**0.8929**
Time(s)	968.04	**174.39**	977.25	3426.15	918.39	948.84	2972.52	1412.13	1209	1271.13

the proposed method, with an OA of 79.00%. However, this model only relies on labeled static receptive fields and a unified information scale within an attention layer, making it unable to simultaneously capture features of different scales. By proposing an SAGC module to address this deficiency, the proposed model produces an OA equal to 81.47%, which is 2.32% higher than S^2ENet. Moreover, it can be noticed that the accuracy is significantly improved in the classification of trees (class 1). This is because the proposed model can more sensitively capture the variation of tree species with respect to their elevation profile, and the combination of HSI and LiDAR can achieve complementary advantages, making it more suitable for tree species identification.

Fig. 6 displays the classification maps of the competitors, the ablation experiments, and the proposed method for the MUUFL Gulfport dataset. Due to irrelevant factors such as noise, many models, such as FusAtNet and Late-Fusion, do not accurately distinguish the boundaries between objects and eventually predict results that do not match the ground truth. The proposed model learns the features from shallow to deep with more adaptable learned features, allowing for a more robust expression of the differences between different

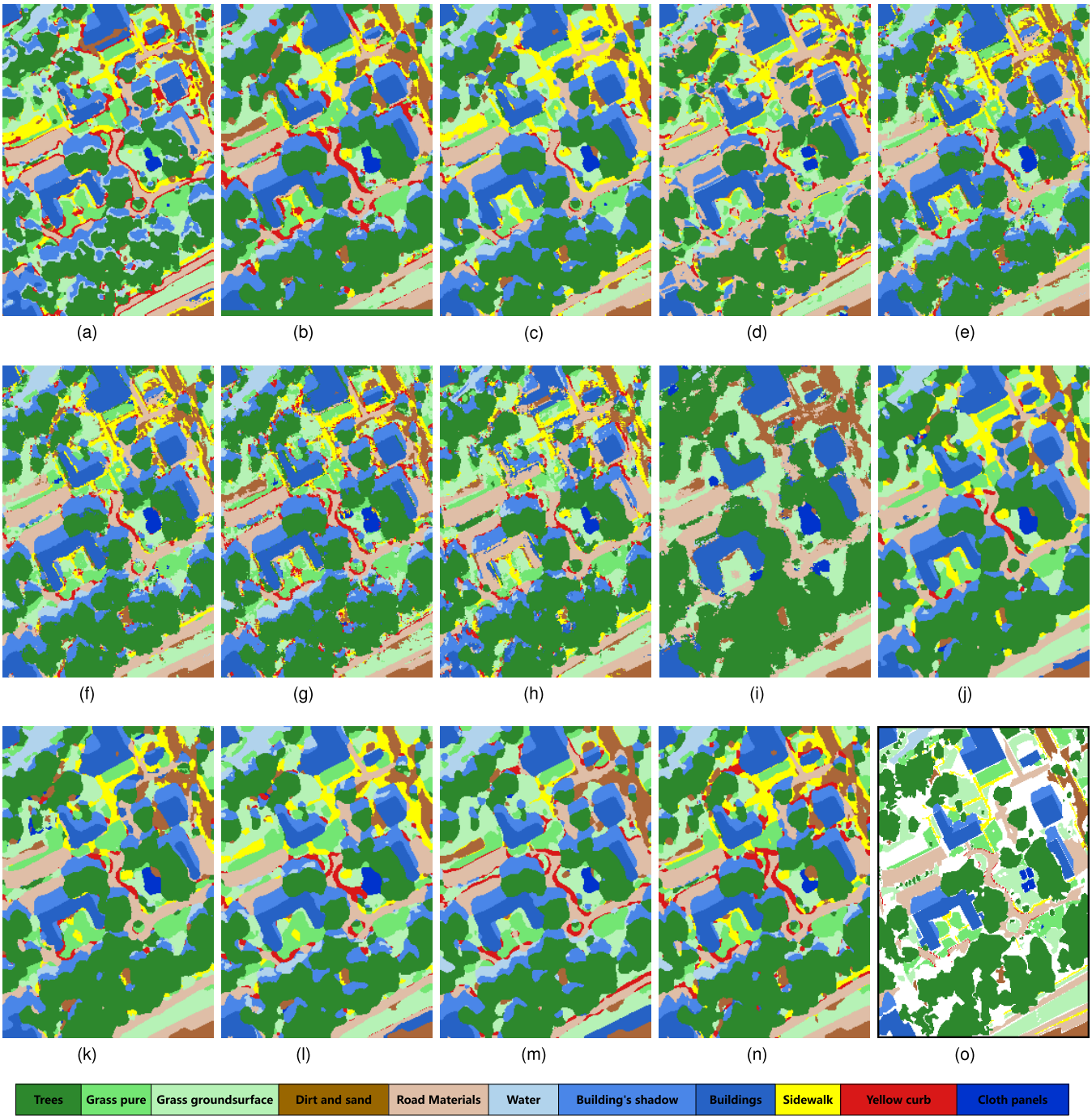

Fig. 6. Classification maps and ground-truth map for the MUUFL Gulfport dataset. (a) CNN-his (73.84%). (b) CoupledCNNs (77.43%). (c) S^2ENet (79.05%). (d) FusAtNet (70.76%). (e) Early-fusion (77.77%). (f) Middle-fusion (78.06%). (g) Late-fusion (76.65%). (h) Cross-HL (73.69%). (i) CALC (77.80%). (j) w/o SEWC (81.01%). (k) w/o SAGC (80.59%). (l) $SAGC_{\_1G}$ (80.80%). (m) $SAGC_{\_2G}$ (81.11%). (n) PSENet (81.47%). (o) Ground-truth map.

categories and smoother visual effects in predicted category maps.

Based on the above analysis, it is evident that the proposed network has the highest OA on the three evaluation metrics in the joint classification of HSI and LiDAR, outperforming the other state-of-the-art competitors. The running times of the proposed method and the competitors for the three datasets are presented in Tables IV–VI, respectively. Generally, the FusAtNet took the longest time in the classification, and CoupleCNNs needed the shortest time. The proposed method obtained the best classification accuracy with an acceptable running time.

E. Classification Results of Ablation Experiments

To verify the significance of the contributions entangled in the proposed network, we present the results of ablation experiments in this section. Table VII shows the numerical values for the four different configurations corresponding to the SAGC and SEWC modules, which are believed to be the main contribution of the proposed method. These four

TABLE VI
CLASSIFICATION ACCURACY FOR THE MUUFL GULFPORT DATASET

Class No.	CNN-HSI	CoupledCNNs	S^2ENet	FusAtNet	Early-Fuison	Middle-Fuison	Late-Fuison	Cross-HL	CALC	PSENet
1	72.78	81.49	78.68	73.49	80.16	79.66	80.52	82.41	**86.39**	85.61
2	**80.98**	68.44	73.80	54.00	64.23	64.24	64.77	66.41	68.54	71.68
3	67.69	56.59	**69.49**	51.76	66.94	65.97	64.57	59.90	50.07	61.94
4	71.26	**89.10**	80.06	65.12	73.40	80.62	73.06	72.06	69.66	88.58
5	75.99	78.61	83.30	83.80	82.92	83.31	78.32	80.31	**87.21**	84.96
6	96.74	**100.00**	99.91	98.25	94.45	99.82	99.65	68.16	99.34	99.98
7	**94.53**	89.35	94.39	72.69	92.15	92.46	90.55	80.31	75.48	87.94
8	73.75	87.89	89.51	85.39	86.33	88.29	83.16	59.49	92.18	**92.51**
9	**51.74**	47.41	52.95	39.39	47.69	49.80	45.59	46.91	0.00	50.49
10	**79.84**	65.32	50.40	46.18	57.51	56.53	60.12	52.37	24.28	60.06
11	73.61	76.25	73.82	74.98	86.22	85.91	85.83	67.95	**93.82**	75.83
OA(%)	73.84	77.43	79.05	70.76	77.77	78.06	76.65	73.69	77.80	**81.47**
AA(%)	76.26	76.40	76.93	67.73	75.64	76.97	75.10	66.93	67.91	**78.14**
Kappa	0.6748	0.7133	0.7356	0.6337	0.7168	0.7217	0.7027	0.6641	0.6791	**0.7624**
Time(s)	1307.22	**215.97**	1503.30	10522.23	1806.66	1650.54	2203.50	2531.13	3354	2352.51

TABLE VII
ABLATION STUDIES ON THREE DATASETS

SZUTree					
Metric	w/o SEWC	w/o SAGC	$SAGC_{\_1G}$	$SAGC_{\_2G}$	PSENet
OA(%)	93.16	93.57	92.97	93.50	**96.91**
AA(%)	87.89	90.37	86.66	91.75	**94.64**
Kappa	0.9016	0.9078	0.8990	0.9070	**0.9555**
Houston2013					
Metric	w/o SEWC	w/o SAGC	$SAGC_{\_1G}$	$SAGC_{\_2G}$	PSENet
OA(%)	89.22	88.65	88.72	89.45	**90.09**
AA(%)	90.48	90.14	90.15	90.71	**91.31**
Kappa	0.8834	0.8773	0.8780	0.8860	**0.8929**
MUUFL Gulfport					
Metric	w/o SEWC	w/o SAGC	$SAGC_{\_1G}$	$SAGC_{\_2G}$	PSENet
OA(%)	81.01	80.59	80.80	81.11	**81.47**
AA(%)	77.97	76.62	75.01	77.31	**78.14**
Kappa	0.7560	0.7510	0.7530	0.7580	**0.7624**

configurations are a network without the SEWC module, a network without the SAGC module, and networks with both SEWC and SAGC but only divided into one group or only two groups when the SAGC module is used. We use "w/o SEWC," "w/o SAGC," $SAGC_{\_1G}$, and $SAGC_{\_2G}$ to represent these four configurations, respectively.

For the network without SAGC, the feature directly enters the SEWC module after feature extraction. At this stage, the number of layers in the network is shallow, the spectral details may not be fully captured, and the spatial distribution of the ground object coverage category may not be effectively represented. The network without SEWC is also unable to emphasize feature categories well. As to our PSENet, the elevation information and spectral information are constrained and fused progressively from shallow to deep, allowing the network to extract more relevant semantic information that can be used for object classification. In addition, as the grouping of SAGC increases, the features extracted and integrated by the network contain more distribution characteristics of the ground objects at different scales, achieving better classification performance. Particularly on the SZUTree dataset, the OA of the proposed method is 3.41% higher than that of the network with only two groups.

We present the visualization results of ablation experiments on three datasets in the subimages (j)–(m) of Figs. 4–6. It can be seen that the proposed method obtains classification result maps with more accurate textures and classification

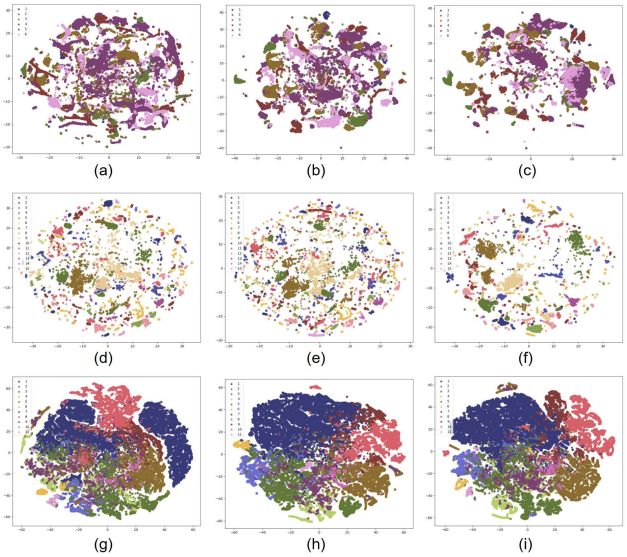

Fig. 7. Feature visualization for different network configurations on three datasets. On the SZUTree dataset: (a) network w/o SEWC, (b) network w/o SAGC, and (c) PSENet. On the Houston2013 dataset: (d) network w/o SEWC, (e) network w/o SAGC, and (f) PSENet. On the MUUFL Gulfport dataset: (g) network w/o SEWC, (h) network w/o SAGC, and (i) PSENet.

effects closer to the ground truth. Using SZUTree as an example, when performing classification using the network with the SAGC of only one group, ficus microcarpa (class 2) is easily misclassified as litchi (class 3). This is likely because ficus microcarpa and litchi both belong to dicotyledonous plants, and it can be challenging to distinguish them from the very similar textures of the leaf. Therefore, when the spatial information or spectral details are not rich enough, it could be difficult for the network to learn discriminative semantic features. To further demonstrate the effectiveness of SAGC and SEWC, we use t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the distribution of features extracted by the network w/o SEWC, w/o SAGC, and the proposed PSENet. Fig. 7 presents the feature distribution in the 2-D space for the three datasets. It can be seen that the feature distribution in Fig. 7 (a), (b), (d), (e), (g), and (h) is more dispersed and disorderly. However, the features learned by the PSENet in Fig. 7 (c), (f), and (i) can improve the

similarity of samples within a class and increase the difference of samples between classes. The results illustrate that the integration of SEWC and SAGC modules in the proposed PSENet is effective in feature extraction for HSI images, which is more beneficial to downstream classification tasks.

V. Conclusion

In this article, we propose an effective network, i.e., PSENet, for the fusion classification of HSI and LiDAR. The network mainly includes two semantic information learning modules to extract semantic information from shallow to deep progressively. SAGC captures the multiscale spatial information of features to fully reflect the texture of ground objects, while SEWC captures the spectral information of features to emphasize the basic properties of ground objects for final classification.

Extensive comparative experiments were performed using three datasets, demonstrating that PSENet can produce excellent overall classification results, outperforming all other state-of-the-art competitors. Ablation experiments were also conducted to verify the contributions of the proposed modules. In summary, the proposed PSENet is a practical deep-learning model for the joint classification of HSI and LiDAR data based on a joint attention mechanism. However, although the proposed PSENet can achieve better classification performance compared with some state-of-the-art methods at small training samples, the potential semantic information of unlabeled samples has not been excavated. It would be helpful to improve the classification performance if the semantic information of unlabeled samples could be considered, which will be on the list of future work.

References

[1] D. Hong et al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[2] L. Zhuang, M. K. Ng, L. Gao, and Z. Wang, "Eigen-CNN: Eigenimages plus eigennoise level maps guided network for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512018.

[3] X. Tong, H. Xie, and Q. Weng, "Urban land cover classification with airborne hyperspectral data: What features to use?" *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 3998–4009, Oct. 2014.

[4] L. Zhuang, X. Fu, M. K. Ng, and J. M. Bioucas-Dias, "Hyperspectral image denoising based on global and nonlocal low-rank factorizations," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10438–10454, Dec. 2021.

[5] M. Rast and T. H. Painter, "Earth observation imaging spectroscopy for terrestrial systems: An overview of its history, techniques, and applications of its missions," *Surv. Geophys.*, vol. 40, no. 3, pp. 303–331, May 2019.

[6] X. Fu, H. Liang, and S. Jia, "Mixed noise-oriented hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5526916.

[7] J. Zhao, Y. Zhong, X. Hu, L. Wei, and L. Zhang, "A robust spectral–spatial approach to identifying heterogeneous crops using remote sensing imagery with high spectral and spatial resolutions," *Remote Sens. Environ.*, vol. 239, Mar. 2020, Art. no. 111605.

[8] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, Aug. 2023.

[9] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[10] M. Hu, C. Wu, and L. Zhang, "GlobalMind: Global multi-head interactive self-attention network for hyperspectral change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 211, pp. 465–483, May 2024.

[11] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, and Q. Li, "Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9553–9568, Nov. 2021.

[12] L. Zhuang, L. Gao, B. Zhang, X. Fu, and J. M. Bioucas-Dias, "Hyperspectral image denoising and anomaly detection based on low-rank and sparse representations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5500117.

[13] X. Fu, Y. Guo, M. Xu, and S. Jia, "Hyperspectral image denoising via robust subspace estimation and group sparsity constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5512716.

[14] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517218.

[15] L.-Z. Huo et al., "Supervised spatial classification of multispectral LiDAR data in urban areas," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0206185.

[16] B. Chen et al., "Multispectral LiDAR point cloud classification: A two-step approach," *Remote Sens.*, vol. 9, no. 4, p. 373, Apr. 2017.

[17] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[18] L. Matikainen, K. Karila, J. Hyyppä, P. Litkey, E. Puttonen, and E. Ahokas, "Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating," *ISPRS J. Photogramm. Remote Sens.*, vol. 128, pp. 298–313, Jun. 2017.

[19] 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.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[21] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[22] P. Ghamisi, R. Souza, J. A. Benediktsson, L. Rittner, R. Lotufo, and X. X. Zhu, "Hyperspectral data classification using extended extinction profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1641–1645, Nov. 2016.

[23] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, Jul. 2017.

[24] M. Salman and S. E. Yüksel, "Fusion of hyperspectral image and LiDAR data and classification using deep convolutional neural networks," in *Proc. 26th Signal Process. Commun. Appl. Conf. (SIU)*, May 2018, pp. 1–4.

[25] D. Xiu, Z. Pan, Y. Wu, and Y. Hu, "MAGE: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539714.

[26] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[27] C. Chen, X. Zhao, W. Li, R. Tao, and Q. Du, "Collaborative classification of hyperspectral and lidar data with information fusion and deep nets," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 2475–2478.

[28] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.

[29] S. K. Roy, A. Sukul, A. Jamali, J. M. Haut, and P. Ghamisi, "Cross hyperspectral and LiDAR attention transformer: An extended self-attention for land use and land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512815.

[30] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, May 2023.

[31] L. Sun, X. Wang, Y. Zheng, Z. Wu, and L. Fu, "Multiscale 3-D–2-D mixed CNN and lightweight attention-free transformer for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 2100116.

[32] M. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[33] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

[34] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[35] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[36] J. Zhang, "Multi-source remote sensing data fusion: Status and trends," *Int. J. Image Data Fusion*, vol. 1, no. 1, pp. 5–24, Mar. 2010.

[37] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 82, pp. 1–18, Jun. 2022.

[38] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[39] C. Ge, Q. Du, W. Li, Y. Li, and W. Sun, "Hyperspectral and LiDAR data classification using kernel collaborative representation based residual fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1963–1973, Jun. 2019.

[40] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[41] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2020.

[42] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

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

[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[46] P. Liu, Y. Ge, L. Duan, W. Li, and F. Lv, "CAFA: Cross-modal attentive feature alignment for cross-domain urban scene segmentation," *IEEE Trans. Ind. Informat.*, vol. 20, no. 10, pp. 11666–11675, Oct. 2024.

[47] R. G. Praveen and J. Alam, "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4803–4813.

[48] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, "Cross-modal fusion and progressive decoding network for RGB-D salient object detection," *Int. J. Comput. Vis.*, vol. 132, no. 8, pp. 3067–3085, Aug. 2024.

[49] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, Jun. 2017.

[50] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[51] C. Li, R. Hang, and B. Rasti, "EMFNet: Enhanced multisource fusion network for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4381–4389, 2021.

[52] Y. Fan et al., "MSLAENet: Multiscale learning and attention enhancement network for fusion classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 10041–10054, 2022.

[53] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based SpectroSpatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 92–93.

[54] S. Fang, K. Li, and Z. Li, "S^{2}ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[55] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "Muufl Gulfport hyperspectral and LiDAR airborne data set," Dept. Elect. Comput. Eng., Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, 2013.

[56] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.

[57] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

Xiyou Fu (Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2012, and the M.S. and Ph.D. degrees from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2015 and 2019, respectively.

He is currently an Assistant Professor with Shenzhen University, Shenzhen, China. His research interests include hyperspectral image restoration, anomaly detection, and super-resolution.

Xi Zhou received the B.E. degree from Jiangxi Normal University, Nanchang, China, in 2021. She is currently pursuing the master's degree in computer science and technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include hyperspectral and light detection and ranging (LiDAR) classification and deep learning.

Yawen Fu received the B.E. degree from Jiangxi Normal University, Nanchang, China, in 2022. She is currently pursuing the master's degree in computer science and technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include remote sensing image classification and carbon stock estimation.

Pan Liu received the B.S. degree from Wuhan Institute of Technology, Wuhan, China, in 2023. He is currently pursuing the M.S. degree with Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and super-resolution.

Sen Jia (Senior Member, IEEE) received the B.S. 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.