frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 09 May 2025
DOI 10.3389/frsen.2025.1545983

Check for updates

OPEN ACCESS

EDITED BY
Qiangqiang Yuan,
Wuhan University, China

REVIEWED BY
Weitao Chen,
China University of Geosciences Wuhan, China
Jing Yao,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Yaohai Lin,
linyaohai@fafu.edu.cn

†PRESENT ADDRESSES
Riqing Chen,
Fujian Key Lab of Agricultural IoT Applications, Sanming University, Sanming, Fujian, China

RECEIVED 16 December 2024
ACCEPTED 23 April 2025
PUBLISHED 09 May 2025

CITATION
Dai M, Liu T, Lin Y, Wang Z, Lin Y, Yang C and Chen R (2025) GLN-LRF: global learning network based on large receptive fields for hyperspectral image classification.
Front. Remote Sens. 6:1545983.
doi: 10.3389/frsen.2025.1545983

COPYRIGHT
© 2025 Dai, Liu, Lin, Wang, Lin, Yang and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLN-LRF: global learning network based on large receptive fields for hyperspectral image classification

Mengyun Dai[1], Tianzhe Liu[2], Youzhuang Lin[1], Zhengyu Wang[3], Yaohai Lin[1]*, Changcai Yang[1] and Riqing Chen[1†]

[1]Fujian Research Institute of Big Data for Agriculture and Forestry, College of Computer andInformation Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [2]Department of Computer Science and Information Security, Police College, Fuzhou, China, [3]Research Institute of Advanced Technology, Wenzhou, China

Deep learning has been widely applied to high-dimensional hyperspectral image classification and has achieved significant improvements in classification accuracy. However, most current hyperspectral image classification networks follow a patch-based learning framework, which divides the entire image into multiple overlapping patches and uses each patch as input to the network. Such locality-based methods have limitations in capturing global contextual information and incur high computational costs due to patch overlap. To alleviate these issues, we propose a global learning network with a large receptive fields network (GLNet) to capture more comprehensive and accurate global contextual information, thereby enriching the underlying feature representation for hyperspectral image classification. The proposed GLNet adopts an encoder-decoder architecture with skip connections. In the encoder phase, we introduce a large receptive field context exploration (LRFC) block to extract multi-scale contextual features. The LRFC block enables the network to enlarge the receptive field and capture more spectral-spatial information. In the decoder phase, to further extract rich semantic information, we propose a multi-scale simple attention (MSA) block, which extracts deep semantic information using multi-scale convolution kernels and fuses the obtained features with SimAM. Specifically, on the IP dataset, GLNet achieved overall accuracies (OA) of 98.72%, average accuracies (AA) of 98.63%, and Kappa coefficients of 98.3%; similar improvements were observed on the PU and HOS18 datasets, confirming its superior performance compared to baseline models. The experimental results demonstrate that GLNet performs exceptionally well in hyperspectral image classification tasks, particularly in capturing global contextual information. Compared to traditional patch-based methods, GLNet not only improves classification accuracy but also reduces computational complexity. Future work will further optimize the model structure, enhance computational efficiency, and explore its application potential in other types of remote sensing data.

KEYWORDS

hyperspectral image classification, multi-scale fusion, spatially separable convolution, large receptive fields, global contextual

1 Introduction

Hyperspectral images provide not only spatial information typically found in natural images but also rich spectral information Chen et al. (2024); Firsov et al. (2024). Each pixel in a hyperspectral image contains dozens or even hundreds of spectral bands. It is the high dimensionality, large amount of data, more spectral information, and high spatial resolution of hyperspectral images that make hyperspectral images more conducive to automatic object identification and classification Zhao D. et al. (2024); Sun et al. (2024). Consequently, hyperspectral image classification has found wide-ranging applications in areas like agricultural monitoring Adão et al. (2017), resource exploration Mohanty et al. (2016), military reconnaissance Tiwari et al. (2011), and urban planning Ghamisi et al. (2014).

In recent years, convolutional neural networks (CNNs) have been widely used to learn spectral-spatial features for hyperspectral image classification. CNNs are capable of extracting multi-dimensional information including spectral, spatial, and spectral-spatial features to achieve improved classification accuracy. The methods used in CNNs for extracting spectral-spatial information typically rely on 2D-CNNs or 3D-CNNs. However, stacked small-scale convolution kernels (e.g., 3×3) used in CNN are sensitive to rotation in high-spectral images. In order to further address the issue of rotation sensitivity in CNN, CASSN Yang K. et al. (2021) proposes a spectral-spatial network with cross-attention to alleviate the impact of image rotation on high-spectral image classification. Subsequently, more researchers have focused on enhancing the rotational invariance robustness of the network. RIAN Zheng et al. (2022) proposes a rotation-invariant attention network, which suppresses redundant spectral information through a central spectral attention module and extracts features through a calibrated spatial attention module. RIAN achieves state-of-the-art performance on various hyperspectral datasets and demonstrates the effectiveness of rotation-invariant attention mechanisms in improving the robustness of deep learning models. State-of-the-art CNN-based hyperspectral image classification methods usually segment images into overlapping small neighborhoods with surrounding pixels and fuse spectral-spatial information using joint statistics and morphological features. However, these methods can only produce shallow appearance features and have insufficient representation ability for high inter-class similarity and large spatial differences, resulting in low classification accuracy. Zheng et al. (2020) propose a fast patch-free global learning (FPGA) framework based on an encoder-decoder fully convolutional network (FCN). They use a global random stratified sampling strategy to obtain different gradients during backpropagation, solving the convergence difficulties in FCN. Moreover, Zhu Q. et al. (2021) propose a spectral-spatial dependent global learning (SSDGL) framework that combines global convolutional long-short term memory (GCL) and global joint attention mechanism (GJAM) to effectively leverage global spatial information. In order to extract deep spectral-spatial features, these methods effectively alleviate the problems of sample scarcity and imbalance through hierarchical balance (H-B) sampling strategies and loss strategies. Compared to CNN-based networks, the aforementioned high-spectral image classification methods based on FCN can learn global information of high-spectral images more effectively, but these methods for extracting spectral-spatial features still rely on small-scale convolution kernels (3×3), making it difficult to obtain context information with a large receptive field and alleviate the problem of image rotation in classification.

Traditional hyperspectral image classification methods typically rely on local feature extraction by analyzing each pixel or small region. While this approach improves classification performance to some extent, it often fails to fully leverage global contextual information when dealing with complex scenarios. Global contextual information is essential for accurate classification, as it provides comprehensive background knowledge, helping to better distinguish between different classes, especially when there are similarities or blurred boundaries between categories. Despite the advances in deep learning for hyperspectral image classification, current methods still face significant challenges, particularly in effectively capturing global contextual information.

Due to the presence of the same object exhibiting different spectral signatures in the collected hyperspectral data, integrating global and contextual information can effectively improve the classification accuracy of hyperspectral images. Existing hyperspectral image classification methods still primarily rely on local feature extraction, such as dividing images into multiple overlapping patches followed by convolution operations. Although this approach can capture local details to some extent, it struggles to model long-range dependencies between distant pixels, resulting in the ineffective utilization of global contextual information. Moreover, convolution kernels with small receptive fields are less adaptable to ground objects with large scale variations, and patch-based processing often introduces excessive redundant computations. Some studies have attempted to enlarge the receptive field using dilated or multi-scale convolutions; however, these approaches may suffer from the loss of fine-grained target information or insufficient fusion of multi-branch features. These limitations become particularly prominent when the number of categories is large, the data is imbalanced, and the spectral dimensionality is high—ultimately constraining classification accuracy and generalization capability. Therefore, developing a method capable of efficiently leveraging large receptive fields for global feature learning has become particularly urgent.

To address the aforementioned problem, we propose a global learning network with large receptive fields (GLNet) based on an encoder-decoder model with skip connections as the basic framework. Specifically, in the context extraction part of the encoder, we propose a large receptive field context exploration (LRFC) block, which combines large convolution kernels, spatially separable convolutions, and dilated convolutions to extract multi-scale context features, enabling effective enlargement of the receptive field while obtaining more spectral-spatial information. To address the issue of further extracting deep semantic information from the encoder, we propose a multi-scale simple attention (MSA) block, which extracts deep semantic information at different scales using multi-scale convolution kernels and fuses the features obtained from different scale branches using SimAM Yang L. et al. (2021).

The main contributions of this paper are as follows:

- Firstly, we propose a novel LRFC block to effectively extract multi-scale contextual features with significant receptive fields on each down-sampling stage of the encoder.

- Secondly, we propose an MSA block to enrich the underlying features representation. MSA block employs attention mechanisms to combine and fuse multi-scale semantic information to further enhance the extraction of deep semantic information in the decoder.
- Thirdly, we propose a novel GLNet based on encoder-decoder architecture to capture more comprehensive and accurate global contextual information for hyperspectral image classification. Our proposed GLNet achieves highly competitive performance compared to several state-of-the-art hyperspectral image classification methods on three benchmark datasets (IP, PU, HOS18).

The rest of this paper is organized as follows. We first review the related Work in Section 2. Then, we describe the details of our proposed GLNet in Section 3. Finally, we give the hyperspectral image classification experimental results in Section 4 and draw the conclusions in Section 5.

2 Related work

2.1 Overview of hyperspectral image classification methods

Hyperspectral images are a valuable source of multidimensional information, but they also suffer from data redundancy, leading to the "curse of dimensionality." In a high-dimensional space, the data density becomes sparse and overfitting can occur when the number of samples is small. In the early stages of hyperspectral image classification research, feature extraction and classifier construction methods are commonly used to overcome this problem. Feature extraction methods project high-dimensional data into a low-dimensional space to reduce the number of features and retain key information, such as principal component analysis (PCA) Licciardi et al. (2012); Prasad and Bruce (2008), independent component analysis (ICA) Villa et al. (2011), and linear discriminant analysis (LDA), *etc.* With the advent of machine learning classifiers, various classifiers, such as support vector machine (SVM) Li et al. (2011), random forest (RF) Breiman (2001); Rodriguez-Galiano et al. (2012), and sparse representation classifiers, are constructed to extract more discriminative spectral information. However, due to the presence of noise, classifiers solely relying on spectral information may not obtain promising classification results. Therefore, researchers have begun to focus on filters to extract more discriminative features. The Gabor filter He L. et al. (2016) and the wavelet filter He et al. (2014) are useful in capturing texture and edge information and analyzing different scale features. However, both methods still have limitations since they do not fully utilize the relationships between pixels, which may lead to incomplete or inaccurate feature extraction. Therefore, morphological methods such as mathematical morphology profile (MP) and extended mathematical morphology profile (EMP) Benediktsson et al. (2005) have been proposed. These methods process the image morphologically and extract the shape and structure features of the image, which can effectively enhance the accuracy and robustness of the image features and make it more suitable for practical applications.

In recent years, deep learning methods have shown great potential in hyperspectral image classification, thanks to various network architectures like Stacked Auto-Encoder (SAE) Chen et al. (2014), Deep Belief Networks (DBN) Chen et al. (2015), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) Makantasis et al. (2015); Guo et al. (2017). These methods can automatically extract features from hyperspectral images, and provide better classification accuracy than traditional machine learning methods. Among them, the CNN-based hyperspectral image classification methods are widely used. They can effectively capture spatial and spectral information in hyperspectral images. Gao et al. Gao et al. (2020) combine t-distributed random neighborhood embedding with CNNs to classify hyperspectral images using 2D-CNNs. Li et al. Li et al. (2020) propose a dual-stream 2D-CNN architecture to better fuse spectral, local, and global spatial features. Some researchers have also employed 3D convolutional kernels to directly extract spectral-spatial information. For example, Zhong et al. Zhong et al. (2017) propose an end-to-end Spectral-Spatial Residual Network (SSRN) based on the combination of spectral and spatial residual blocks, which first extracts spectral features using 3D convolution in the spectral dimension, then uses 3D convolution in the spatial domain to extract spatial features for hyperspectral image classification. Paoletti et al. Paoletti et al. (2018) construct a deep residual network (PResNet) by stacking pyramid bottleneck residual units Han et al. (2017) to extract more complex spatial and spectral features as the network deepens.

Given the superiority of the attention mechanism in the handling of long-range information, the Transformer architecture has generated significant research interest and practical applications in the field of hyperspectral image classification (HSI)Zhang et al. (2022); Roy et al. (2023); Yang X. et al. (2022). In particular, He et al. He et al. (2019) proposed a bidirectional encoder representation transformer network (HSI-BERT), primarily built on a multi-head self-attention (MHSA) mechanism in an MHSA layer. He et al.He et al. (2024) proposed an Interval Group Spatial-Spectral Mamba framework (IGroupSS-Mamba),Benefiting from the Interval Group S6 Mechanis (IGSM),IGroupSS-Mamba achieves non-redundant spatial-spectral dependencies modeling across different feature groups. Ahmad et al. Ahmad et al. (2024)proposed a WaveFormer, which combines the power of wavelet transforms and ViT for hyperspectral image classification. Pang et al.Pang et al. (2025) proposed the Mambahsi model, a spatial–spectral joint processing approach for hyperspectral image classification. By employing a multi-scale convolutional feature extraction module, the model significantly improves classification accuracy, highlighting the importance of integrating spatial and spectral information in complex scenarios. Yao et al.Yao et al. (2023) designed the ExViT framework by combining multimodal learning with a vision Transformer architecture. They integrated multiple remote sensing data sources and applied attention mechanisms to capture long-range dependencies, which enhanced the model's understanding of global contextual information and improved the robustness and accuracy of land use and land cover classification.

However, CNN's reliance on patch-based input makes the central class prediction heavily dependent on the surrounding context, affecting the network's performance in terms of patch size setting. In general, larger patches can capture richer contextual features, leading to better classification performance. However, as the patch size expands, the overlap between adjacent patches similarly increases, resulting in a surge in storage and computational costs for the network. This makes it challenging to achieve a balance between classification accuracy and network efficiency. Furthermore, the final classification in CNN is implemented through fully connected layers, which flatten the feature maps into 1-D vectors, further compromising information. To address these issues, in the image segmentation field, Jonathan et al. proposed Fully Convolutional Networks (FCN) Long et al. (2015), which can take arbitrary-sized feature maps as inputs and produce pixel-level outputs for semantic segmentation tasks. Due to the ability of FCNs to handle inputs of any size and generate predictions for each pixel, outputting as feature maps, it has also been increasingly applied in hyperspectral image classification. F-CNN Li et al. (2018) utilizes the PCA algorithm to extract the first principal component (PC) as the training label. Then, the training data consists of hyperspectral data and copies of the first PC. FCSPN Jiang et al. (2021) integrates a 3-D fully convolution network (3D-FCN) with a convolutional spatial propagation network (CSPN) for HSI classification, effectively reducing computational complexity while enhancing the algorithm's adaptive ability. UML Wang et al. (2022) proposes a multi-scale spatial channel attention mechanism and multi-scale shuffle blocks, considering both effective spectral information and contextual information, improving the redundancy operations and land cover map distortion issues in hyperspectral image classification.

2.2 Attention mechanism

Attention mechanisms are becoming increasingly popular in hyperspectral image classification frameworks. These mechanisms enhance regions with informative data and suppress regions that have minimal impact on classification or contain noise. Researchers have conducted in-depth and extensive studies on attention mechanisms. For instance, Hu et al. Hu et al. (2018) propose SENet, which uses channel-wise attention mechanisms based on squeeze-and-excitation modules. Woo et al. Woo et al. (2018) address the issue of neglecting the importance of spatial features in image classification by SENet, which only focuses on the relationship among features across different channels, and propose a novel attention module that incorporates both channel and spatial attention mechanisms for a more comprehensive approach. Attention mechanisms have great potential in hyperspectral image classification since hyperspectral images have fine internal structures and provide spectral features from a large number of channels, making them suitable for the application of both spatial and channel-wise attention mechanisms. Mei et al. Mei et al. (2019) propose an algorithm that utilizes bidirectional CNNs to extract spectral and spatial features separately, which are then combined in the classification network. Attention mechanisms are introduced during the feature extraction process in both branches, with fully connected layers used to calculate attention weights for spectral and spatial domains. DBMA Ma et al. (2019) proposes a dual-branch multi-attention mechanism network that also incorporates both channel and spatial attention. The dual-attention network based on self-attention mechanism (DANet) Fu et al. (2019) combines local features and global dependencies. The spectral-spatial attention block in RSSAN Zhu MH. et al. (2021), similar to the CBAM Woo et al. (2018), directly operates on the hyperspectral raw image and extracts spectral and spatial features from it. The Central Attention Network (CAN) Liu et al. (2022) employs a dense strategy to extract spectral-spatial information based on the similarity weights obtained from the query pixel and its surrounding pixels.

2.3 Multi-scale feature extraction

Hyperspectral image classification is commonly regarded as a multi-classification task in high-altitude remote sensing, which requires the extraction of features from objects of different sizes, making multi-scale fusion increasingly significant in computer vision and remote sensing. Furthermore, the process of hyperspectral image classification involves classifying datasets that have a significant difference in size. Hence, it is essential to design a multi-scale fusion framework that can classify objects with varying sizes to improve classification performance. MSSN Wu et al. (2019) employs a dual-branch structure to jointly extract spectral and spatial features in a multi-scale spectral-spatial domain. However, MSSN separately extracts spectral and spatial information in the dual-branch structure, neglecting the interaction between spectral and spatial information. Pooja et al. Pooja et al. (2019) propose a multi-scale residual convolutional neural network (MSR-CNN) that combines multi-scale, extended convolutional kernels, and residual connections based on the CNN framework. Nevertheless, the multi-scale feature extraction module in MSR-CNN fails to consider the importance of different features in classification performance during information fusion. CSMS-SSRN Lu et al. (2020) introduces a three-branch structure, where each branch extracts spectral and spatial features of different sizes, followed by multi-scale fusion. However, this approach only extracts spectral features of one size and spatial features of another size from a single branch, resulting in insufficient feature fusion. Subsequently, MSF-MIF Yang L. et al. (2022) proposes a more comprehensive fusion approach based on CSMS-SSRN that achieves multi-scale fusion at the spectral level and fully considers the fusion of different scale features from a spatial perspective.

3 Proposed method

Figure 1 illustrates the pipeline of the proposed GLNet, including the feature encoder module and the feature decoder module. In the following, we first introduce the basic network structure of the proposed GLNet in Section 3.1. Then, we describe the details of the proposed LRFC block and MSA block in Section 3.2 and Section 3.3, respectively.

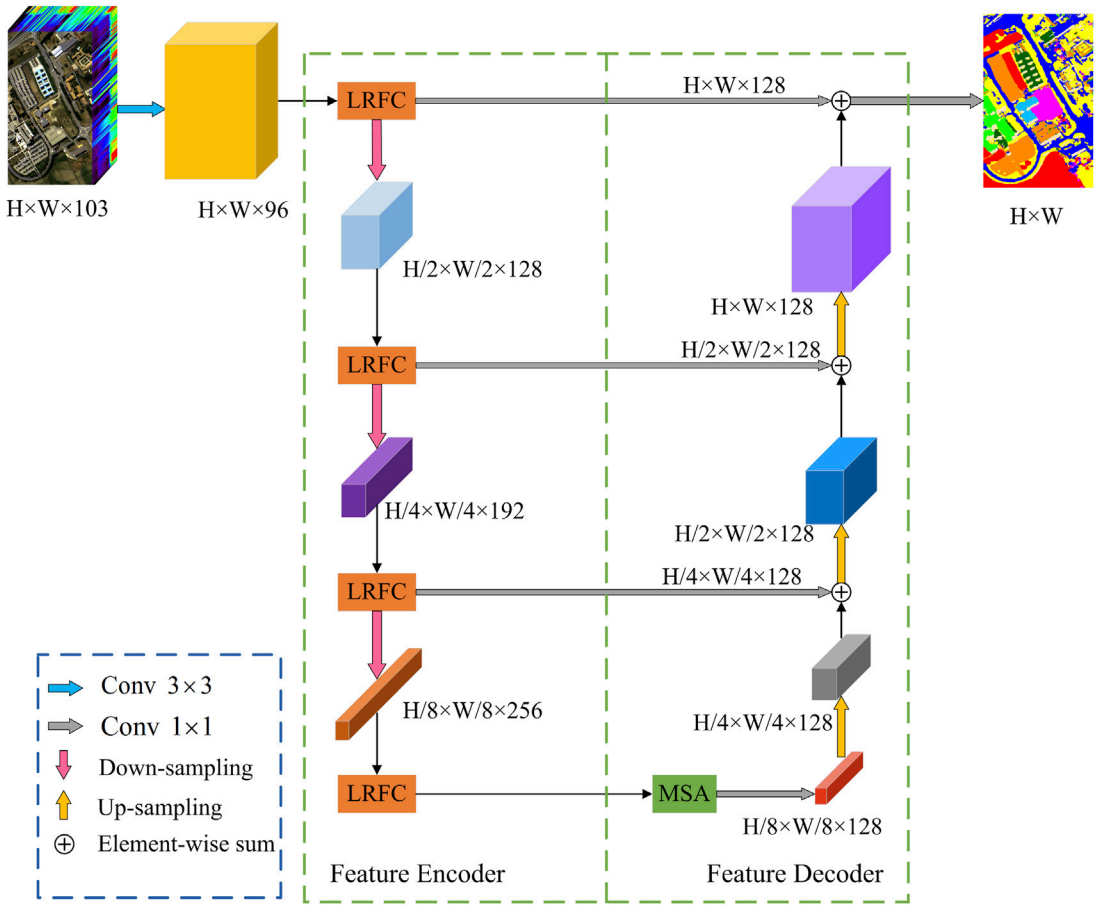

FIGURE 1
The pipeline of the proposed GLNet. The proposed GLNet primarily comprises two parts: the encoder and the decoder. The encoder includes the down-sampling and LRFC blocks, while the decoder consists of the MSA block and up-sampling.

3.1 Overview of GLNet

Inspired by the global learning strategies of FPGA [50] and SSDGL [51], we propose a novel GLNet for high-dimensional image classification, as shown in Figure 1, which utilizes a fully convolutional encoder-decoder network as the basic structure, and combines shallow feature information captured through effective large receptive fields with multi-scale deep feature extraction techniques to improve feature representation. In addition, it considers the spectral relationships between different bands and the spatial correlations among individual pixels to enhance classification performance.

For each down-sampling section in the encoder, we adopt a basic structure consisting of 3 × 3 convolutional layers, followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation layer. While 3 × 3 convolutional layers are commonly used for feature extraction in deep convolutional networks, they are sensitive to rotation in high-spectral images. In order to extract robust spectral and spatial features, we propose the large receptive field context (LRFC) block, which adopts multi-level large-scale convolution kernels and incorporates spatially separable convolution and dilated convolution to balance performance and efficiency. As a result, the LRFC block can obtain a larger spectral-spatial receptive field and extract richer contextual information.

In the decoder, we propose a multi-scale simple attention (MSA) block to further extract three-dimensional information of spectral and spatial features by using multi-scale convolution kernels and SimAM. This enables rich extraction of deep features for hyperspectral images. For the up-sampling section, a transposed convolution operation of 3 × 3 is used in each up-sampling, followed by a group normalization (GN) layer and a ReLU activation layer, similar to the down-sampling section. Furthermore, the deep features obtained by the hyperspectral image after passing through the encoder serve as the input of the bottom layer. The features upsampled from the previous layer in the decoder are fused with the features from each layer of the encoder through skip connections to obtain enhanced features as the input of the next layer in the decoder. This fusion method combines strong semantic information with more detailed spatial information. Finally, the features are gradually restored to the original spatial size of the hyperspectral image, and a classification probability map of the same spatial size as the input image is output.

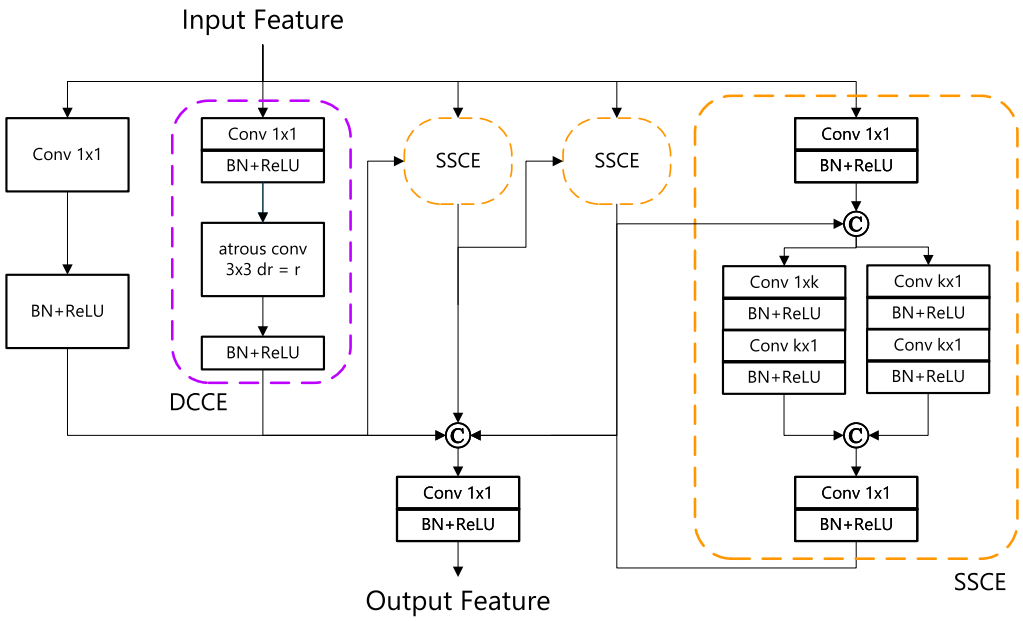

FIGURE 2
The structure diagram of LRFC block. The proposed LRFC block mainly consists of five branches, including the spatially separated convolution for context exploration (SSCE) module and the dilated convolution for context exploration (DCCE) module.

3.2 LRFC block

To obtain multi-level contextual information, three main methods have been widely used, including using large convolutional kernels, stacking small convolutional kernels, and using dilated convolutions. The core idea of these three methods is to expand the receptive field to extract contextual information. Among these methods, stacking multiple layers of small convolutional kernels as done in Unet Ronneberger et al. (2015) and ResNet He KM. et al. (2016), is a common way to extract contextual information. Using large convolutional kernels allows for directly obtaining a large receptive field, but it can increase computational complexity and memory consumption, requiring certain device requirements. Dilated convolution is a sparse sampling method, but it can also lead to information loss of small-scale objects due to the varying sizes of objects in remote sensing images. Using a larger dilation rate to obtain a larger field of view can also result in the loss of contextual information of small-scale objects.

To address the challenge of balancing efficiency and performance in contextual information extraction, this paper proposes the large receptive field contextual (LRFC) block to expand the field of view while maintaining computational efficiency and memory utilization and avoiding the problem of losing information of small-scale objects. The LRFC module significantly enlarges the receptive field of the network by integrating three strategies—large-kernel convolution, dilated convolution, and spatially separable convolution—without substantially increasing computational cost. Specifically, large-kernel convolutions directly cover broader regions, enabling the capture of long-range dependencies between distant pixels. Dilated convolutions further allow the network to sample more contextual information across multiple scales. Meanwhile, spatially separable convolutions offer a balance between a large receptive field and reduced computational overhead.

Through a multi-branch structure, the module performs parallel extraction of multi-scale features and subsequently fuses them, achieving a balance between local detail preservation and global structural representation. This design effectively enhances the expression of spectral–spatial collaborative features in hyperspectral images, enabling the network to better capture global contextual information. Consequently, it allows for more accurate discrimination between categories that are spectrally similar but spatially distinct. As shown in Figure 2, the LRFC block consists of two parts, the spatially separable convolutional context exploration (SSCE) module and the dilated convolutional context exploration (DCCE) module.

In the LRFC block, the upper part comprises a branch structure that preserves information from the upper layer and four cascaded branch structures with different scales of kernels. The branch structure that preserves information from the upper layer includes a 1×1 convolutional layer, a BN layer, and a ReLU activation function. The DCCE module is used in the minimum scale branch of the four cascaded branch structures to extract small-scale features in the image. In the other three large-scale branch structures, the SSCE module is used to expand the field of view and improve feature extraction performance. To combine information from different receptive fields, information flow is added between adjacent SSCE blocks, further expanding the effective receptive field of the SSCE module to capture information across the entire feature region. The output information from the current block is then fused with the original feature information and used as input for the next block. This method can share features of different scales and enable the output information of the current SSCE module to be better

utilized by the next SSCE module. As a result, it improves the ability to perceive a wider context compared to simple parallel branch structures.

To enable the context extraction block to have a larger receptive field while being efficient and effective, we adopt the SSCE block, shown as the yellow dashed box in Figure 2. The SSCE block utilizes multi-scale large kernels combined with spatially separable convolutions to extract context information. First, for a given input feature, a $k = 1$ convolution layer is used to reduce the number of channels. Then, we use two parallel spatially separable convolutions $\{1 \times k_i, k_i \times 1\}$ and $\{k_i \times 1, 1 \times k_i\}$ to extract context features. In this method, we use a large-scale kernel $(k_1 = 27, k_2 = 29, k_3 = 31)$ to further increase the effective receptive field when combined with spatially separable convolutions, making the context information extracted more rich and diverse. We adopt parallel spatially separable convolutions to approximate the feature extraction effect of standard convolutions, achieving a better balance between performance and efficiency. When a standard convolution layer and a spatially separable convolution have the same input tensor shape $(W \times H \times C_{in})$ and output tensor shape $(W \times H \times C_{out})$ and use the same convolution kernel size $(k \times k)$, the number of parameters and computations of the standard convolution are as Formulas 1, 2:

$$P_{\text{std}} = k^2 \times C_{in} \times C_{out} \tag{1}$$

$$C_{\text{std}} = k^2 \times W \times H \times C_{in} \times C_{out} \tag{2}$$

The number of parameters and computational cost of spatially separable convolution are as Formulas 3, 4 respectively

$$P_{\text{ssc}} = 2 \times k \times C_{in} + C_{in} \times C_{out} \tag{3}$$

$$C_{\text{ssc}} = 2 \times k \times W \times H \times C_{in} + W \times H \times C_{in} \times C_{out} \tag{4}$$

From the equation above, we can see that, for standard convolution, the kernel size k and the number of parameters have a quadratic relationship, while for spatially separable convolution, it exhibits a simple linear relationship. Therefore, when extracting features using large-scale kernels, the difference in parameter and computational complexity between regular convolutional and spatially separable convolutional operations is substantial. In other words, spatially separable convolutional operations are more efficient while maintaining comparable classification performance.

After obtaining the features from two parallel spatially separable convolutions, we combine them using a $k = 1$ convolutional layer to restore the original channel dimension. To capture non-linear features in the contextual information more effectively, we include BN and ReLU activation functions between each convolutional layer. BN helps normalize the input data, making the network easier to train and decreasing the risk of overfitting. ReLU activation functions can introduce non-linearities into the network and improve its feature extraction ability.

In order to increase the receptive field of small-scale convolution operations, we propose the DCCE block, as shown in the purple dashed box in Figure 2, which leverages dilated convolutions to extract contextual features from small-scale information. Specifically, compared with the SSCE block, the DCCE block performs a relatively simple operation, which is to apply dilated convolution on the given input information. Given that the DCCE block mainly focuses on small-scale information, we set the value of k to three and the dilation rate to 3. The down-sampling operation using dilated convolution can preserve more detailed information and enlarge the effective receptive field, which is beneficial for processing local features and fine details, and ultimately improve the learning ability of the network. It is worth noting that a dilated convolutional layer with a dilation rate can achieve a receptive field comparable to that of a standard convolutional layer with smaller parameter sizes (Formula 5)

$$\eta = (k - 1) \times r + 1, \tag{5}$$

where η represents the receptive field, and k and r represent the kernel size and dilation rate, respectively. After the dilated convolution operation, BN and ReLU activation function are also used for non-linear operation.

3.3 MSA block

To further extract deep information, we propose a multi-scale simple attention (MSA) block in the decoder, as shown in Figure 3. The MSA module further enhances the acquisition of global contextual information by combining multi-scale convolutions with lightweight attention mechanisms. Specifically, parallel convolutional kernels of different scales are employed to capture multi-level features ranging from fine-grained local details to global structural patterns. Subsequently, lightweight attention mechanisms such as SimAM are used to assign point-wise adaptive weights, emphasizing pixel positions that are more relevant to discrimination. By adopting this "multi-scale feature extraction + adaptive attention" strategy, the MSA module effectively integrates deep semantic information with shallow details, while fully leveraging both the spectral and spatial dimensions of hyperspectral imagery. This facilitates precise modeling of global contextual relationships in complex scenes. The MSA block is employed to extract deep abstract semantic information. As the network deepens, the overlapping areas between receptive fields increase, which leads to the acquisition of more coarse-grained abstract information and global image information.

The MSA block employs a multi-scale triplet branching structure to improve the network's ability to extract high-resolution information from hyperspectral remote sensing images, as shown in the diagram. To extract feature information at different levels of abstraction, multi-scale convolution structures with $k = 3, 5$, and seven are used in this block. BN layers and ReLU activation functions are employed after the convolution layers to obtain non-linear feature information. In the fusion part, in addition to using 1×1 convolution layers to transform channels, we also introduce the SimAM to better fuse different multi-scale information.

SimAM is an attention module based on neuroscience theory that exploits the importance of each neuron by optimizing an energy function. Unlike 1-D or 2-D weight attention modules, which treat every neuron in a channel or spatial position equally and may have limitations in learning distinctive features. SimAM estimates the importance of individual neurons based on the energy function and generates three-dimensional weights while achieving a lightweight implementation.

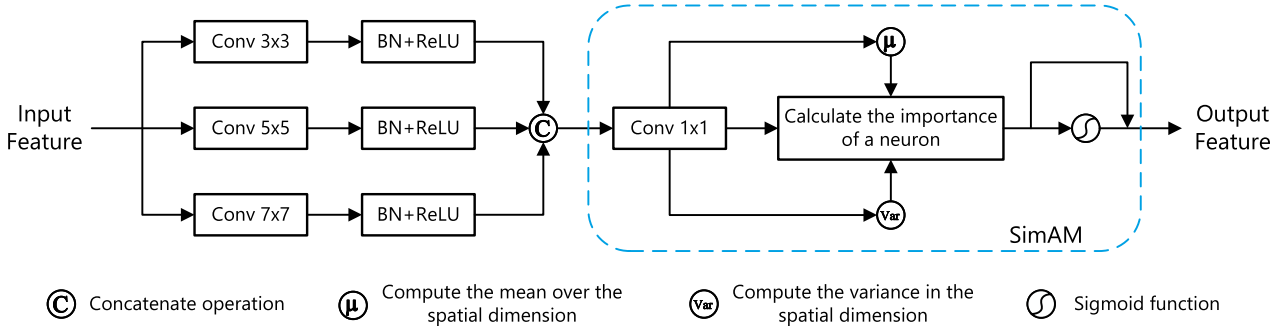

FIGURE 3
The structure diagram of MSA block.

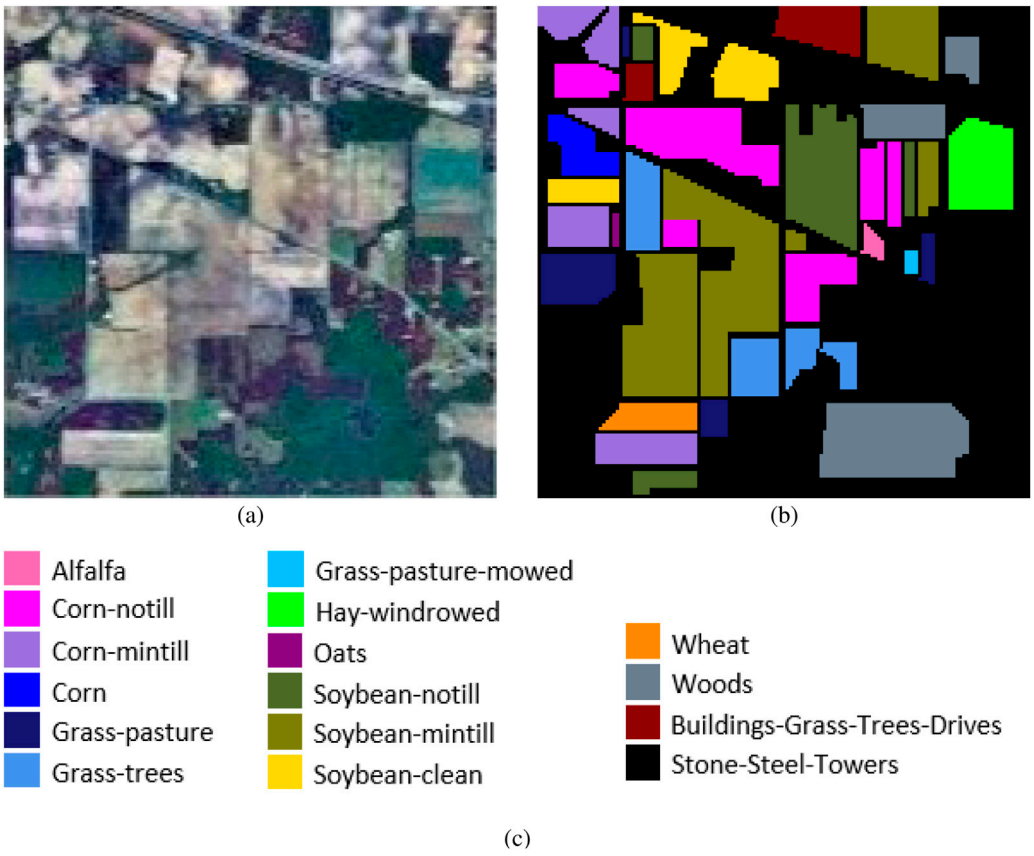

FIGURE 4
IP data visualization. **(a)** Three-band false color composite. **(b)** Ground-truth map **(c)** Legend.

SimAM generates effective three-dimensional weights by estimating the importance of each neuron. A smaller energy function value indicates greater differences between the neuron and its surrounding neurons, resulting in stronger discriminability and richer information content, both of which are crucial in visual processing. Consequently, the importance of a single neuron is determined as Formula 6:

$$E_i = \frac{(x_i - \mu)^2}{4(\sigma^2 + \lambda)} + \frac{1}{2}, \tag{6}$$

where x_i is the input feature value of the neuron, μ and σ^2 are the mean and variance of all input feature values of neurons in the channel, respectively. λ is a smoothing term used to avoid division by a variance less than zero, thus ensuring the stability of the energy function calculation. All neurons in the channel share the same mean and variance. The first term $\frac{(x_i-\mu)^2}{4(\sigma^2+\lambda)}$ in the equation represents the degree of difference between the neuron and other neurons in the channel. A smaller energy function value for a neuron, in comparison to other neurons, implies less similarity, which indicates greater importance. The second term $\frac{1}{2}$ is a constant

TABLE 1 The number of training and testing samples in the IP dataset.

No.	Class	Train	Test	Total
1	Alfalfa	5	41	46
2	Corn-notill	72	1356	1428
3	Corn-mintill	42	788	830
4	Corn	12	225	237
5	Grass-pasture	25	458	483
6	Grass-trees	37	693	730
7	Grass-pasture-mowed	5	23	28
8	Hay-windrowed	24	454	478
9	Oats	5	15	20
10	Soybean-notill	49	923	972
11	Soybean-mintill	123	2,332	2,455
12	Soybean-clean	30	563	593
13	Wheat	11	194	205
14	Woods	64	1201	1265
15	Building-grass-trees-drives	20	366	386
16	Stone-steal-towers	5	88	93
Total		529	9,720	10249

bias that does not affect the value of the energy function but is crucial for gradient calculations.

Finally, the output of the SimAM can be represented as the dot product of the input feature X and the importance of each neuron E_i, expressed as Formula 7:

$$Y = X \times \sigma(E_i), \tag{7}$$

where $\sigma(\cdot)$ represents the activation function used to map the importance of each neuron to a value between 0 and 1.

4 Experimental results

In this section, we first introduce the dataset, evaluation indicators, and experimental parameter settings in Section 4.1, Then, we present quantitative comparison and qualitative analysis in Section 4.2. Finally, we show the results of ablation experiments conducted under different modules in Section 4.3.

4.1 Data description and evaluation index introduction

IndianPines (IP): This is a hyperspectral image dataset acquired by the AVIRIS airborne sensor over northwest Indiana in the United States in 1992. The spectral range of the spectrometer is 400–2500nm, and after removing the zero and water absorption bands, a total of 200 bands were used for classification. The dataset includes 16 classes, 10249 labeled pixels, and a spatial size of 145 × 145 pixels with a spatial resolution of 20mpp. The visualization of this dataset is shown in Figure 4 and Table 1 provides the names of each class, the division of the training and testing data, and the number of labeled samples in each class.

Pavia University (PU):This is a hyperspectral image dataset captured by the ROSIS sensor in 2002, covering the Pavia region in the north of Italy. It contains a large amount of plant category data and it mainly includes urban image data such as roads, buildings, and urban landscape vegetation, with a total of nine categories and 42,776 labeled pixels. After removing the noisy bands, there are still 103 bands left, and the dataset has a spatial resolution and size of 1.3mpp. The visualization of this dataset is shown in Figure 5 and Table 2 provides the name of each category, the split of training and test sets, and the number of labeled samples. DFC Houston 2018 (HOS18): It is an open high-spectral-image dataset consisting of 48 bands with a resolution of 1 m. Additionally, it includes three bands of LiDAR data with a resolution of 0.5 m and high-resolution image data with a resolution of 0.05 m. The dataset contains 20 categories and was first introduced in the 2018 IEEE GRSS Data Fusion Contest[1]. It has been used for research in high-spectral-image classification. The visualization of this dataset is shown in Figure 6 and Table 3 provides the name of each category, the split of training and test sets, and the number of labeled samples.

To quantitatively evaluate the classification performance of different models, we adopt the overall accuracy (OA), average accuracy (AA), per-class accuracy, and Kappa coefficient as evaluation metrics for model classification.

4.2 Comparisons with state-of-the-art methods

To ensure the accuracy and reproducibility of our experiments, we employ high-performance hardware and state-of-the-art software frameworks. Our experimental platform consists of a 12th generation Intel(R) Core(TM) i9-12900K processor with 16 cores, 24 threads, 12M cache, and a processing speed of 3.19 GHz, as well as an NVIDIA GeForce RTX 3090 graphics card with 24G VRAM. For the development environment, we select Python 3.8 and conduct experiments using the PyTorch framework. We adopt the parameter settings in SSDGL and train the IP and PU datasets for 600 epochs and the HOS18 dataset for 1000 epochs, respectively. For optimization, we use SGD, with the momentum set to 0.9, weight decay set to 0.001, and an initial learning rate of 0.005, which is multiplied by $(1 - \frac{\text{iter}}{\text{max\_iter}})^{\text{power}}$ with the iteration number, where power = 0.8 and max_iter = 1000.

We compare GLNet with nine high-spectral image classification methods, including M3D-CNN He et al. (2017), 3DDLA-CNN Hamida et al. (2018), PResNet Paoletti et al. (2019), MS^{3}A-Net Dai et al. (2022), DBSSAN Zhao J. et al. (2024), MTMSD Zhou et al. (2024),U^{2}ConvFormer Zhan et al. (2024) FPGA Zheng et al. (2020), and SSDGL Zhu Q. et al. (2021),which is our benchmark model. We conduct a detailed analysis of the classification performance of the proposed GLNet and employ three evaluation metrics (OA, AA, and Kappa coefficient) to quantify its classification performance. The

1 https://hyperspectral.ee.uh.edu/2018IEEEDocs/DataReport.pdf

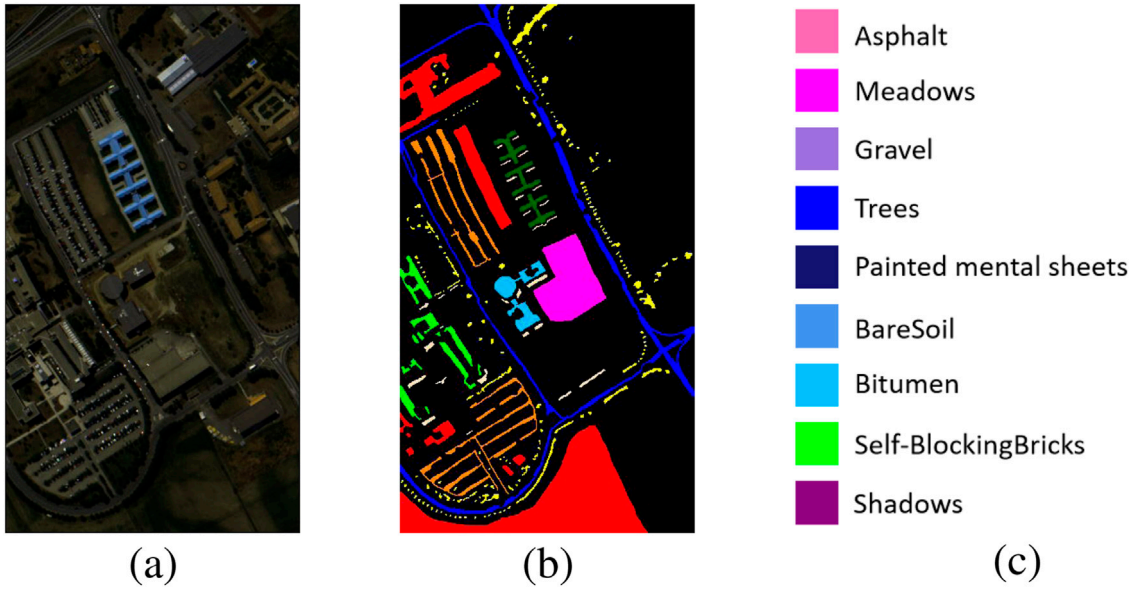

FIGURE 5
PU data visualization. **(a)** Three-band false color composite **(b)** Ground-truth map **(c)** Legend.

TABLE 2 The number of training and testing samples in the PU dataset.

No.	Class	Train	Test	Total
1	Asphalt	67	6,564	6,631
2	Meadows	187	18462	18649
3	Gravel	21	2078	2099
4	Trees	31	3,033	3,064
5	Painted metal sheets	14	1331	1345
6	Bare soil	51	4,978	5,029
7	Bitumen	14	1316	1330
8	Self-blocking bricks	37	3,645	3,682
9	Shadows	10	937	947
Total		432	42344	42776

best results are highlighted in bold. The benchmark model is underlined.

The classification performance of different methods on the IP dataset is shown in Table 4. FCN-based methods achieve better classification accuracy, with an overall accuracy (OA) of over 90%. Among the 16 categories, the proposed GLNet achieves the highest accuracy, which can be attributed to its use of global learning that leverages global spatial context information at the bottom level. GLNet achieves an improvement of 2%–5% compared to FPGA and 1%–2% compared to SSDGL. In particular, GLNet shows a 5%–7% improvement in corn classification compared to FPGA and a slight improvement over SSDGL. The performance enhancement of GLNet can be primarily attributed to the LRFC block, which provides a larger receptive field and richer contextual feature information. The LRFC block enables the model to better capture both local and global information, particularly in a limited dataset. Additionally, its hierarchical balanced sampling strategy helps mitigate the issues posed by insufficient and imbalanced data. Consequently, GLNet achieves robust classification performance even when the dataset is small or imbalanced. As shown in Table 4, FCN-based methods also yield higher accuracy in OA, AA, and Kappa coefficients. GLNet outperforms several state-of-the-art methods in terms of overall accuracy (OA), average accuracy (AA), and Kappa coefficient. Specifically, for land cover types with blurred boundaries and similar spectral characteristics—such as Grass-pasture vs Grass-trees and Soybean-notill vs Soybean-min till—traditional methods often suffer from misclassification. In contrast, GLNet effectively captures long-range contextual information through its large receptive field module, significantly improving classification accuracy for these challenging categories. Moreover, for classes with very limited training samples, such as Alfalfa, GLNet leverages its multi-scale attention mechanism to better extract discriminative features, maintaining a high recognition rate. These results demonstrate that GLNet exhibits greater robustness and discriminative capability when dealing with typical challenges in the IP dataset, including small sample sizes, spectrally similar classes, and complex scene structures.

The PU dataset comprises a substantial number of samples and exhibits exceptional spatial resolution, making spatial information crucial for discerning difficult-to-classify categories in hyperspectral imagery. As depicted in Table 5, among the nine categories, the proposed GLNet achieves the highest accuracy. This is because the integration of an LRFC block in GLNet enables the acquisition of global spatial context information and the extraction of interdependencies between spectral channels. Notably, the FCN-based approach significantly outperforms the CNN-based approach in the categories of “Gravel”, “Bitumen”, and “Self-blocking bricks”. The proposed GLNet achieves the highest accuracy in terms of OA, AA, and kappa coefficient. Compared to MS^{3}A-Net, the proposed method achieves an improvement of approximately 5% in

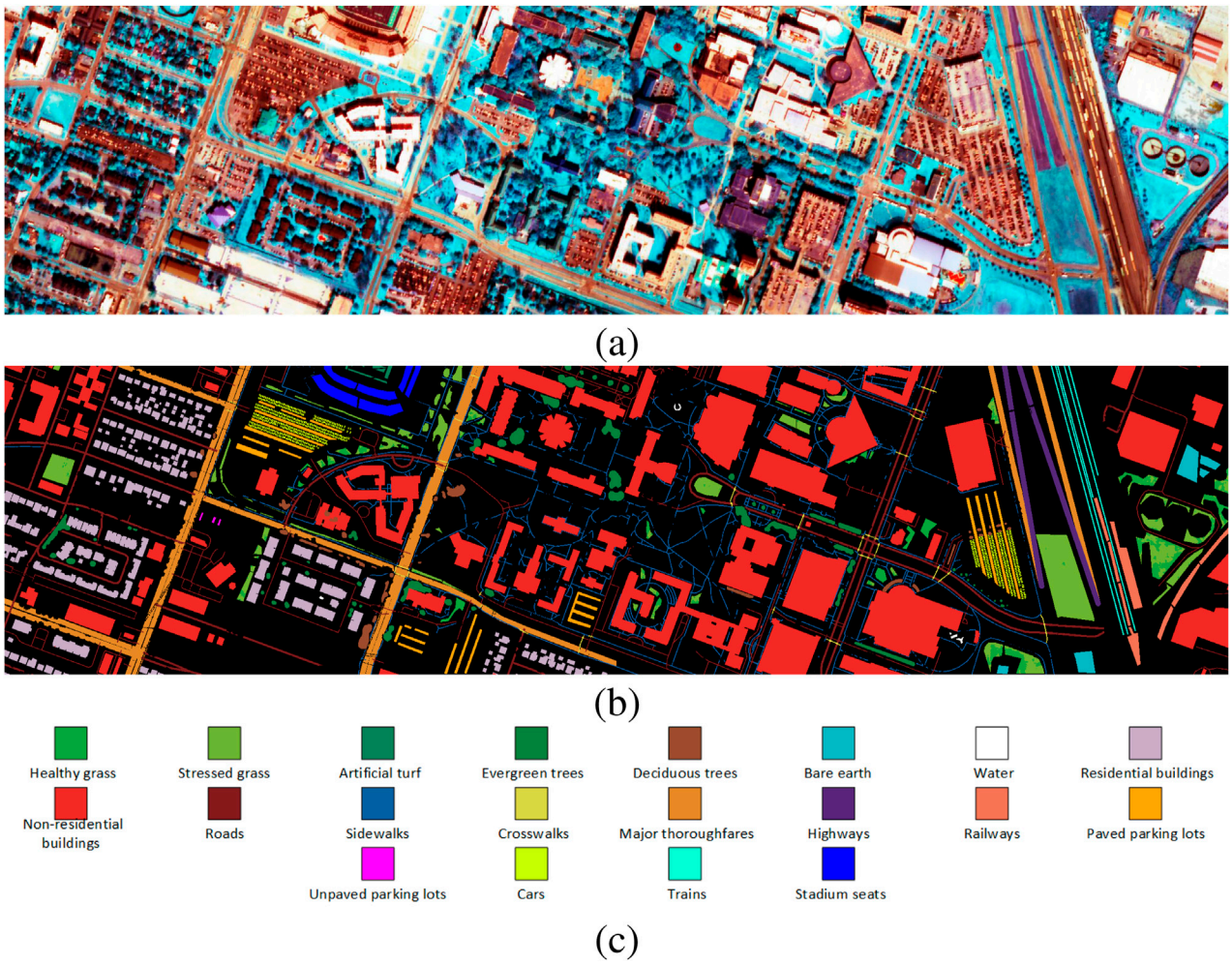

FIGURE 6
HOS18 data visualization. **(a)** Three-band false color composite **(b)** Ground-truth map **(c)** Legend.

classification performance, and an improvement of approximately 1% compared to the FPGA method. Furthermore, the proposed method demonstrates improved classification performance compared to SSDGL. In particular, for categories such as Asphalt, Gravel, and Bitumen, which exhibit high spectral similarity and are often misclassified by traditional models, GLNet significantly improves class separability and reduces confusion by incorporating multi-scale attention mechanisms and global context modeling. Moreover, for small-area classes with irregular boundaries—such as Self-Blocking Bricks and Shadows—GLNet leverages its multi-scale feature fusion capability to more accurately restore spatial details and boundary information, thereby enhancing classification completeness and continuity. Overall, the results demonstrate that GLNet exhibits stronger adaptability and robustness in handling the challenges of the PU dataset.

The HOS18 dataset has a large spatial scale and an uneven distribution of samples across different categories. For instance, while the "Non-residential buildings" category has 223,752 samples, the "Unpaved parking lots" category contains only 146 samples. CNN-based networks, which primarily learn local features, typically require a larger number of training samples compared to FCN-based networks. However, in HOS18, only 5% of the samples are selected as training data for CNN-based networks, while FCN-based networks use only 10 samples per category for training, leaving the rest for testing. As shown in Table 6, methods like M3D-CNN and SSDGL perform poorly in the "Unpaved parking lot" category, while FPGA shows relatively low accuracy in the "Crosswalks" category. Although methods like 3DDLA-CNN, PResNet, and MS^{3}A-Net outperform the aforementioned methods in certain categories, they still struggle with accurate classification in the "Crosswalks" category. In contrast, the proposed GLN-LRF excels in all 20 categories, achieving high accuracy across the board. This remarkable performance can be attributed to the LRFC and MSA blocks, which are key components of GLN-LRF. These blocks address the challenge of imbalanced samples, particularly in categories with fewer samples, by capturing rich spatial and spectral context. The LRFC block enables the model to capture larger receptive fields, improving the classification of both small and large categories. Additionally, the MSA block enhances the model's ability to integrate multi-scale semantic information, further improving its performance on imbalanced datasets. Overall,

TABLE 3 The number of training and testing samples in the HOS18 dataset.

No.	Class	Train	Test	Total
1	Healthy grass	10	9,789	9,799
2	Stressed grass	10	32492	32502
3	Artificial turf	10	674	684
4	Evergreen trees	10	13585	13595
5	Deciduous trees	10	5,011	5,021
6	Bare earth	10	4,506	4,516
7	Water	10	256	266
8	Residential buildings	10	39762	39772
9	Non-residential buildings	10	223742	223752
10	Roads	10	45856	45866
11	Sidewalks	10	34019	34029
12	Crosswalks	10	1508	1518
13	Major thoroughfares	10	46338	46348
14	Highways	10	9,855	9,865
15	Railways	10	6,927	6,937
16	Paved parking lots	10	11490	11500
17	Unpaved parking lots	10	136	146
18	Cars	10	6,537	6,547
19	Trains	10	5,359	5,369
20	Stadium seats	10	6,814	6,824
Total		200	504656	504856

GLN-LRF achieves the highest accuracy in terms of OA, AA, and Kappa coefficient among the seven methods, with scores of 98.72%, 98.63%, and 98.3%, respectively. These results highlight GLN-LRF's superior ability to handle imbalanced datasets and achieve robust classification across diverse categories.

To ensure a fair comparison among different classification methods, we randomly select 5% of the samples as the training set and the remaining samples as the test set for all classification methods in the HOS18 dataset. The comparison results are shown in Table 7. As shown in the table, it can be observed that the number of training samples has little effect on the FCN-based classification methods. The proposed GLNet method outperforms the other methods in terms of different training sample sizes and achieves the highest scores in the three classification evaluation metrics.

To further qualitatively analyze the classification performance of different hyperspectral image classification methods, we compare the classification results of M3D-CNN, 3DDLA-CNN, PResNet, MS^{3}A-Net, FPGA, SSDGL, and GLNet on the IP, PU, and HOS18 datasets using visualization graphs. Figure 7 shows the classification results visualization of different classification methods on the IP dataset. It can be observed that the FCN-based classification method exhibits superior visual performance compared to the CNN-based method, with smoother images and fewer noisy points. This is because the FCN-based method can effectively leverage global contextual information to obtain a complete surface cover structure, extract the most discriminative spatial features, and attain category boundaries closer to the real image. Compared to CNN-based methods, FCN-based methods perform better in classifying categories with similar features, such as corn and soybeans, and possess better generalization ability. Our proposed GLNet outperforms the FPGA method in the classification of the "Soybean-mintill" and "Soybean-notill" categories.

The visualization of the classification results of different hyperspectral image classification methods on the PU dataset is shown in Figure 8. It can be observed that both CNN-based and FCN-based classification methods perform well on the PU dataset, owing to its high spatial resolution of 1.3mpp and large sample size. In particular, based on the classification results of the "Meadows" class in the middle of the image, it can be seen that the FCN-based exhibits advantages and does not exhibit salt-and-pepper noise. This is likely due to the fact that the grassland area in the middle of the image is relatively large, and FCN is a global learning method that can effectively utilize surrounding information. Additionally, the LRFC block proposed in this paper has a larger receptive field, which allows it to obtain more surrounding information, and the MSA block can better extract deep semantic information. Therefore, the proposed GLNet in this paper can obtain clear class boundaries and complete road structures. The visualization of the classification results using different classification methods on the HOS18 dataset is shown in Figure 9, indicating significant visual differences between the methods. A zoomed-in box in the upper right corner of the figure exposes notable misclassification in the "Non-residential buildings" category for the CNN-based classification method. Notably, although the MS^{3}A-Net method is a CNN-based method, it demonstrates evident improvement compared to other CNN-based methods, including M3D-CNN, 3DDLA-CNN, and PResNet, and even performs better in classifying deciduous trees and roads in the zoomed-in box, compared to the two FCN-based classification methods, FPGA and SSDGL. The proposed GLNet yields the best visual effects on the classification of "Non-residential buildings," "Deciduous trees," and "Roads" compared to other methods, with clear boundaries between different categories and no obvious noise points within the same category. Furthermore, the "Sidewalks" category is more detailed, and the proposed GLNet also exhibits good performance in this category, which is likely because the LRFC block has a large receptive field that can explore long-distance contextual information, and the MSA block has the ability to extract and integrate deep semantic information. The results suggest that GLNet outperforms other methods in terms of fine details and boundary clarity on the HOS18 dataset.

4.3 Ablation studies

In this section, we conduct ablation studies on the blocks and comparative experiments on classification performance using different convolution kernels to further verify the effectiveness of LRFC and MSA blocks in GLNet. The training and test set splits for

TABLE 4 Comparison of classification accuracy (%) and Kappa measure of different classification methods on the IP dataset (5% training set).

Classes	CNN-based						Transformer-based	FCN-based		
	M3D-CNN	3DDLA-CNN	PResNet	MS^{3}A-Net	DBSSAN	MTMSD	U^{2}ConvFormer	FPGA	SSDGL	GLNet
1	40.48	45.24	46.51	44.19	**100.00**	**100.00**	**100.00**	92.68	**100.00**	**100.00**
2	72.53	64.99	86.54	91.65	97.73	99.52	98.57	98.82	99.78	**99.93**
3	**52.10**	41.10	91.98	91.72	99.06	99.01	97.28	94.54	99.49	**99.87**
4	52.29	40.83	77.27	67.73	97.75	99.62	99.69	**100.00**	**100.00**	**100.00**
5	86.26	78.15	91.78	82.67	**100.00**	98.62	99.08	83.62	99.78	**100.00**
6	99.11	96.13	96.17	96.76	**100.00**	99.97	99.54	97.98	**100.00**	**100.00**
7	30.77	3.85	65.38	15.38	**100.00**	99.20	94.67	**100.00**	**100.00**	**100.00**
8	99.77	99.55	93.48	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
9	44.44	27.78	73.68	89.47	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
10	82.44	47.65	87.40	91.93	99.53	98.79	99.58	94.15	99.78	**100.00**
11	56.71	72.64	93.48	95.71	99.00	99.67	**99.82**	95.97	98.54	99.74
12	61.36	25.64	80.62	86.41	94.09	98.84	98.32	97.87	99.29	**99.64**
13	**100.00**	90.48	**100.00**	94.24	**100.00**	99.78	97.84	96.91	**100.00**	**100.00**
14	97.08	95.36	93.80	97.79	**100.00**	99.86	99.76	99.67	**100.00**	**100.00**
15	54.37	42.82	94.44	84.72	**100.00**	99.42	99.14	99.73	**100.00**	**100.00**
16	60.00	89.41	95.35	94.19	92.41	98.10	95.64	**100.00**	**100.00**	**100.00**
OA	73.19	68.18	90.71	92.28	98.75	99.45	99.16	96.69	99.51	**99.90**
AA	68.11	60.10	85.49	82.79	98.01	99.40	98.68	97.00	99.79	**99.95**
Kappa	69.75	63.48	89.41	91.19	98.66	99.37	99.04	96.23	99.44	**99.88**

TABLE 5 Comparison of classification accuracy (%) and Kappa measure of different classification methods on the PU dataset (1% training set).

Classes	CNN-based						Transformer-based	FCN-based		
	M3D-CNN	3DDLA-CNN	PResNet	MS^{3}A-Net	DBSSAN	MTMSD	U^{2}ConvFormer	FPGA	SSDGL	GLNet
1	93.82	92.11	96.33	90.67	99.78	**100.00**	99.92	99.58	**100.00**	**100.00**
2	90.31	89.85	96.72	98.75	99.93	**100.00**	99.97	99.87	99.99	**100.00**
3	48.12	2.55	69.61	84.00	**100.00**	**100.00**	99.92	99.95	**100.00**	**100.00**
4	89.12	72.50	97.21	95.56	99.72	99.59	98.47	99.27	99.34	**100.00**
5	95.72	99.55	**100.00**	99.62	**100.00**	**100.00**	99.92	**100.00**	**100.00**	**100.00**
6	43.28	28.32	87.15	97.46	99.56	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
7	62.41	0.00	67.54	84.74	**100.00**	**100.00**	99.52	**100.00**	**100.00**	**100.00**
8	80.99	85.68	80.4	93.42	99.66	99.87	99.53	99.83	**100.00**	**100.00**
9	97.12	0.00	97.6	100	**100.00**	99.79	98.56	**100.00**	**100.00**	**100.00**
OA	81.82	72.60	92.05	95.55	99.83	99.95	99.76	99.81	99.95	**100.00**
AA	77.88	52.28	88.06	93.80	99.75	99.92	99.53	99.83	99.93	**100.00**
Kappa	75.66	62.50	89.45	94.11	99.78	99.94	99.70	99.74	99.93	**100.00**

TABLE 6 Comparison of classification accuracy (%) and Kappa measure of different classification methods on HOS18 dataset. The training set is 5% for CNN-based networks and 10 samples for each class for FCN-based networks.

Classes	CNN-based						Transformer-based	FCN-based		
	M3D-CNN	3DDLA-CNN	PResNet	MS^{3}A-Net	DBSSAN	MTMSD	U^{2}ConvFormer	FPGA	SSDGL	GLNet
1	89.06	89.85	83.53	88.00	**90.73**	88.87	87.26	50.75	70.75	89.88
2	95.26	94.97	96.20	**96.81**	95.63	96.29	90.11	96.08	91.53	95.32
3	82.62	82.93	**100.00**	**100.00**	**100.00**	99.93	99.33	16.48	**100.00**	**100.00**
4	96.35	97.29	98.83	98.51	94.66	99.32	98.87	90.34	96.27	**99.72**
5	74.38	91.53	95.04	94.59	90.86	97.01	95.74	0.00	12.93	**99.54**
6	93.22	98.14	99.90	99.52	97.83	**100.00**	99.97	0.62	93.92	**100.00**
7	95.65	88.54	97.98	**100.00**	**100.00**	97.40	99.87	75.39	99.61	**100.00**
8	85.40	89.37	97.80	97.23	96.64	99.81	99.75	93.63	97.57	**99.98**
9	94.62	93.58	99.12	99.12	96.64	99.72	99.53	98.55	99.20	**99.95**
10	61.44	76.00	86.74	87.20	90.6	96.63	94.92	61.32	69.14	**97.21**
11	64.12	68.10	82.30	81.02	78.84	93.96	90.16	37.75	51.00	**95.17**
12	0.00	15.11	24.98	23.00	26.65	55.30	52.94	0.00	0.13	**98.41**
13	77.11	74.43	94.72	95.17	84.99	97.87	97.85	76.29	86.81	**98.36**
14	75.22	92.82	98.86	98.33	91.8	99.27	99.45	64.77	96.38	**99.99**
15	98.13	99.92	**100.00**	99.60	98.64	99.92	99.98	70.26	96.78	**100.00**
16	89.16	97.38	97.62	97.21	95.91	**99.84**	99.45	62.39	87.83	99.76
17	0.00	95.68	60.29	86.76	98.65	**100.00**	97.60	0.00	**100.00**	**100.00**
18	78.12	83.76	96.60	96.18	92.17	**99.60**	98.92	59.59	73.96	99.28
19	92.63	90.49	99.98	99.66	98.58	**99.99**	99.93	83.30	87.16	99.96
20	92.15	96.53	99.12	99.81	99.58	**100.00**	99.97	87.11	99.74	**100.00**
OA	85.93	88.09	95.54	95.57	92.94	98.29	97.70	81.96	88.84	**98.72**
AA	76.73	85.78	90.48	91.89	90.56	96.04	95.38	55.95	80.54	**98.63**
Kappa	81.71	84.64	94.19	94.23	90.70	97.77	97.01	76.26	85.35	**98.33**

the IP, PU, and HOS18 datasets are shown in Tables 1–3, respectively.

To investigate the effectiveness of large-scale spatially separable convolution kernels in expanding the effective receptive field, we varied the spatially separable scales in the SSCE module to [3, 5, 7], [9, 11, 13], [15, 17, 19], [21, 23, 25], and [27, 29, 31], while maintaining all other conditions constant. The classification results on the HOS18, IP, and PU datasets with different kernel scales are shown in Figure 10. As the IP dataset has a smaller spatial scale, there is no significant difference between the large and small scales of the kernel. However, the classification accuracy achieved using the large-scale kernel was nearly identical to that of the small-scale kernel, with an OA value exceeding 99.9%, indicating outstanding classification performance. The HOS18 dataset has a larger spatial size, which makes the large-scale kernel more advantageous compared to the small-scale kernel. This advantage becomes more pronounced as the kernel size increases. Thus, the results indicate that large-scale kernels can capture information from a larger range, which is beneficial for high-dimensional hyperspectral image classification tasks, particularly for hyperspectral data with large spatial dimensions. This approach can further improve the classification performance of the network.

We added the LRFC block after each down-sampling operation in the encoder and used the MSA block in the decoder of the GLNet. To analyze and verify the effectiveness of these two blocks in the network, we designed an experiment. Specifically, we compared three cases using the basic encoder-decoder structure as the baseline. The first case was to only add the LRFC block in the encoder's down-sampling operation, the second case was to only add the MSA block in the decoder, and the third case was to add both LRFC and MSA blocks in the encoder and decoder (referred to as GLNet). Table 8 shows the three classification metrics (OA, AA, and Kappa) on the

TABLE 7 Comparison of classification accuracy (%) and Kappa measure of different classification methods on the HOS18 dataset (5% training set).

Classes	CNN-based						Transformer-based	FCN-based		
	M3D-CNN	3DDLA-CNN	PResNet	MS^{3}A-Net	DBSSAN	MTMSD	U^{2}ConvFormer	FPGA	SSDGL	GLNet
1	89.06	89.85	83.53	88.00	**90.73**	88.87	87.26	50.65	73.32	88.80
2	95.26	94.97	96.20	**96.81**	95.63	96.29	96.08	90.07	88.80	95.23
3	82.62	82.93	**100.00**	**100.00**	**100.00**	99.93	99.33	16.02	**100.00**	**100.00**
4	96.35	97.29	98.83	98.51	94.66	99.32	98.87	90.37	93.70	**99.88**
5	74.38	91.53	95.04	94.59	90.86	97.01	95.74	0.00	60.35	**99.39**
6	93.22	98.14	99.90	99.52	97.83	**100.00**	99.97	0.61	99.32	**100.00**
7	95.65	88.54	97.98	**100.00**	**100.00**	97.40	99.87	74.60	98.41	**100.00**
8	85.40	89.37	97.80	97.23	96.64	99.81	99.75	93.64	98.64	**99.98**
9	94.62	93.58	99.12	99.12	96.64	99.72	98.53	98.55	**99.57**	**99.96**
10	61.44	76.00	86.74	87.20	90.6	96.63	94.92	61.29	80.71	**96.94**
11	64.12	68.10	82.30	81.02	78.84	**93.96**	90.16	37.80	43.95	89.87
12	0.00	15.11	24.98	23.00	26.65	55.30	52.94	0.00	0.00	**98.68**
13	77.11	74.43	94.72	95.17	84.99	97.87	97.85	76.25	95.87	**98.36**
14	75.22	92.82	98.86	98.33	91.8	99.27	99.45	64.70	99.18	**99.99**
15	98.13	99.92	**100.00**	99.60	98.64	99.92	99.98	70.49	96.45	**100.00**
16	89.16	97.38	97.62	97.21	95.91	**99.84**	99.45	62.43	92.45	99.76
17	0.00	95.68	60.29	86.76	98.65	**100.00**	97.60	0.00	**100.00**	**100.00**
18	78.12	83.76	96.60	96.18	92.17	**99.60**	98.92	59.91	80.54	99.41
19	92.63	90.49	99.98	99.66	98.58	**99.99**	99.93	83.33	89.73	99.96
20	92.15	96.53	99.12	99.81	99.58	**100.00**	99.97	87.07	99.86	99.98
OA	85.93	88.09	95.54	95.57	92.94	98.29	97.70	81.95	91.03	**98.32**
AA	76.73	85.78	90.48	91.89	90.56	96.04	95.38	55.89	84.50	**98.31**
Kappa	81.71	84.64	94.19	94.23	90.70	97.77	97.01	76.24	88.24	**97.81**

HOS18 dataset under different block combinations. From Table 8, we can clearly observe that adding either LRFC or MSA block in a single subnetwork, compared to the baseline, resulted in significant improvements in all three metrics. Compared with the baseline, the LRFC subnetwork improved all three metrics by 10%, with an increase of 19% in AA metric, and the MSA subnetwork improved all three metrics by approximately 6%–12%. Furthermore, the GLNet network, which included both LRFC and MSA blocks, showed significant improvements in all classification metrics compared to the baseline. This indicates that the LRFC and MSA blocks are effective in extracting contextual features in hyperspectral image classification and that their combination can complement each other, further improving the network's classification performance.

To further validate the effectiveness of different blocks, we compare the feature maps generated by the ablation studies. Figure 11 clearly shows the differences among them. It is evident that both LRFC subnetwork and MSA subnetwork display significant improvements over Baseline, particularly for the "Cars" and "Paved parking lots" categories in the gray box on the left side of the figure. Baseline's feature maps produce messy and disordered classification results for these two categories. On the contrary, the "Cars" and "Paved parking lots" feature maps generated by the LRFC and MSA subnetworks are neatly arranged, yielding well-defined classification results. Furthermore, the feature maps of the proposed GLNet also exhibit significant changes compared to the other three feature maps, as illustrated in the gray box on the right of Figure 11, where "Trains" and "Railways" are represented as straight and well-defined borders with their surrounding classes. Overall, the addition of the LRFC and MSA blocks significantly mitigates noise points in classification maps, making the boundaries between different categories more distinct and recognizable. This highlights the effectiveness of these two blocks to extract contextual features from hyperspectral images in a complementary manner, enhancing the classification performance of the network.

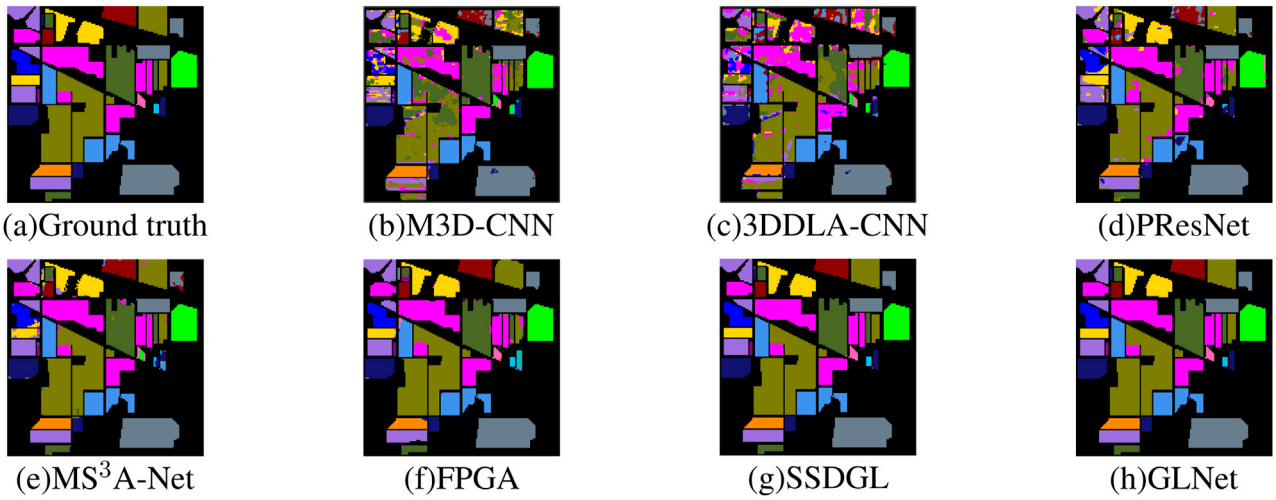

FIGURE 7
Color classification maps of four different algorithms on the IP dataset. **(a)** Ground truth **(b)** M3D-CNN **(c)** 3DDLA-CNN **(d)** PResNet **(e)** MS^{3}A-Net **(f)** FPGA **(g)** SSDGL **(h)** GLNet.

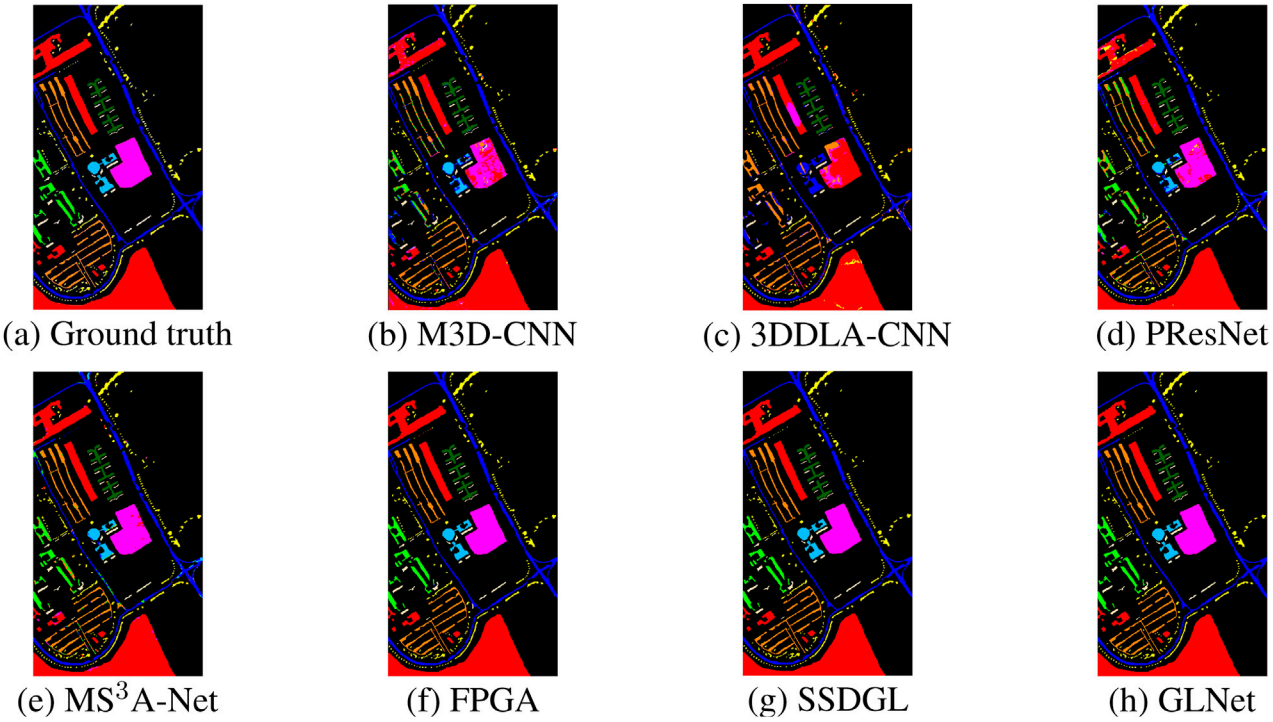

FIGURE 8
Color classification maps of four different algorithms on the PU data. **(a)** Ground truth **(b)** M3D-CNN **(c)** 3DDLA-CNN **(d)** PResNet **(e)** MS^{3}A-Net **(f)** FPGA **(g)** SSDGL **(h)** GLNet.

5 Conclusion

This paper alleviates the challenge of CNN-based classification methods failing to learn global spectral-spatial information. We propose a GLNet based on an encoder-decoder network structure. Firstly, in the encoder, LRFC block is proposed to extract contextual information of a large receptive field in the spectral space. Secondly, in the decoder, a new MSA block is proposed to further extract deep semantic information, enhancing the feature learning capability of the network. In the experimental section, we compare various state-of-the-art classification methods quantitatively and qualitatively on three commonly used datasets. The results demonstrate that the

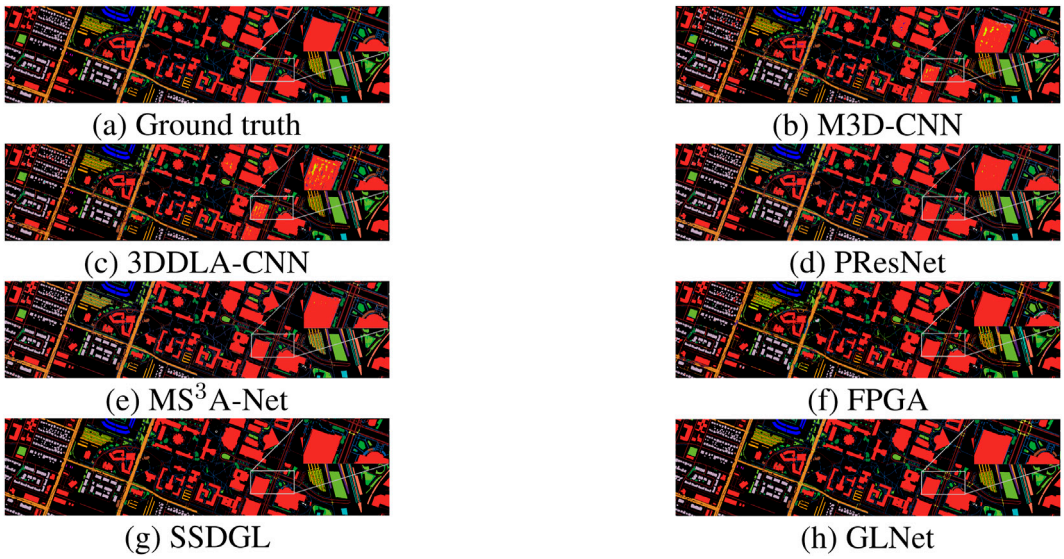

FIGURE 9
Color classification maps of four different algorithms on the HOS18 dataset. **(a)** Ground truth **(b)** M3D-CNN **(c)** 3DDLA-CNN **(d)** PResNet **(e)** MS^{3}A-Net **(f)** FPGA **(g)** SSDGL **(h)** GLNet.

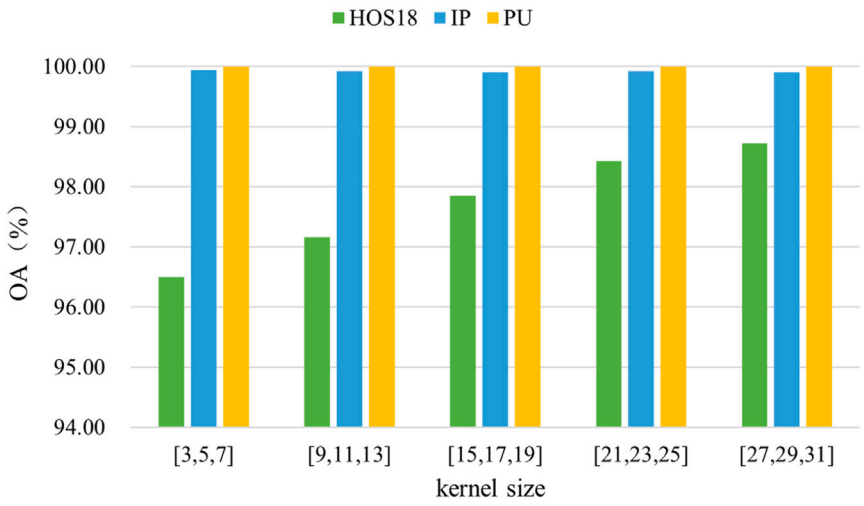

FIGURE 10
The impact of different scales of convolution kernels in the LRFC block on OA.

TABLE 8 Classification results of different blocks in GLNet under HOS18 dataset.

Module	LRFC	MSA	HOS18		
			OA	AA	Kappa
Baseline	×	×	88.39	79.42	84.76
LRFC	✓	×	98.50	98.43	98.05
MSA	×	✓	94.39	91.80	92.67
GLNet	✓	✓	**98.72**	**98.63**	**98.33**

proposed GLNet has a competitive advantage in hyperspectrkal image classification. The GLNet model proposed in this study demonstrates excellent performance in hyperspectral image classification; however, it has some limitations. First, its high computational complexity and reliance on high-performance hardware may restrict its applicability in real-time or resource-constrained environments. Second, while the model shows strong results on the datasets tested, its performance is still dependent on the dataset's characteristics, and it may not perform well with small sample sizes or in cases of class imbalance. Furthermore, the complexity of the deep learning architecture reduces its

FIGURE 11
Color classification maps of four different algorithms on the IP dataset. **(a)** Baseline **(b)** LRFC **(c)** MSA **(d)** GLNet.

interpretability, which could pose challenges in practical applications. Future work should focus on improving computational efficiency, enhancing the model's generalization across diverse datasets, and exploring integration with other advanced models. Additionally, addressing challenges such as small sample learning and transfer learning will help expand the model's applicability and further boost its performance.

Overall, the proposed GLNet model demonstrates promising performance in improving the classification accuracy of hyperspectral images. Future research may further explore the application of GLNet to other types of remote sensing data, such as multispectral imagery and synthetic aperture radar (SAR) data, in order to evaluate its generalizability and effectiveness across different data modalities. Additionally, attention should be given to assessing the robustness of GLNet under varying environmental conditions, including changes in illumination, seasonal variations, and geographical diversity, which may affect classification accuracy. These investigations will not only expand the potential applications of GLNet but also further validate its effectiveness and reliability in complex and diverse real-world scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

MD: Writing – original draft, Writing – review and editing. TL: Writing – review and editing. YoL: Writing – review and editing. ZW: Writing – review and editing. YaL: Writing – review and editing. CY: Writing – original draft, Writing – review and editing. RC: Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work is supported in part by the Science and Technology Innovation Guiding Project of Agricultural Science and Technology Autonomous Innovation Funds of Ningxia Academy of Agricultural and Forestry Sciences under Grant NKYG-25-16, and in part by the Science and Technology Innovation Special Fund Project of FAFU under Grant KFB24043.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., et al. (2017). Hyperspectral imaging: a review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sens.* 9, 1110. doi:10.3390/rs9111110

Ahmad, M., Ghous, U., Usama, M., and Mazzara, M. (2024). Waveformer: spectral–spatial wavelet transformer for hyperspectral image classification. *IEEE Geoscience Remote Sens. Lett.* 21, 1–5. doi:10.1109/lgrs.2024.3353909

Benediktsson, J. A., Palmason, J. A., and Sveinsson, J. R. (2005). Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geoscience Remote Sens.* 43, 480–491. doi:10.1109/tgrs.2004.842478

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, F., Su, B., and Jia, Z. (2024). Tuh-nas: a triple-unit nas network for hyperspectral image classification. *Sensors* 24, 7834. doi:10.3390/s24237834

Chen, Y., Lin, Z., Zhao, X., Wang, G., and Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 7, 2094–2107. doi:10.1109/jstars.2014.2329330

Chen, Y., Zhao, X., and Jia, X. (2015). Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 8, 2381–2392. doi:10.1109/jstars.2015.2388577

Dai, M., Sun, Q., Dai, L., Lin, Y., Wei, L., Yang, C., et al. (2022). Ms3a-net: multi-scale and spectral-spatial attention network for hyperspectral image classification. *Int. J. Remote Sens.* 43, 7139–7160. doi:10.1080/01431161.2022.2155081

Firsov, N., Myasnikov, E., Lobanov, V., Khabibullin, R., Kazanskiy, N., Khonina, S., et al. (2024). Hyperkan: Kolmogorov–arnold networks make hyperspectral image classifiers smarter. *Sensors* 24, 7683. doi:10.3390/s24237683

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Gao, L. R., Gu, D. X., Zhuang, L. N., Ren, J. C., Yang, D., and Zhang, B. (2020). Combining t-distributed stochastic neighbor embedding with convolutional neural networks for hyperspectral image classification. *IEEE Geoscience Remote Sens. Lett.* 17, 1368–1372. doi:10.1109/lgrs.2019.2945122

Ghamisi, P., Dalla Mura, M., and Benediktsson, J. A. (2014). A survey on spectral–spatial classification techniques based on attribute profiles. *IEEE Trans. Geoscience Remote Sens.* 53, 2335–2353. doi:10.1109/tgrs.2014.2358934

Guo, T., Dong, J., Li, H., and Gao, Y. (2017). "Simple convolutional neural network on image classification," in *International conference on big data analysis (ICBDA)*, 721–724.

Hamida, A. B., Benoit, A., Lambert, P., and Amar, C. B. (2018). 3-d deep learning approach for remote sensing image classification. *IEEE Trans. Geoscience Remote Sens.* 56, 4420–4434. doi:10.1109/tgrs.2018.2818945

Han, D., Kim, J., and Kim, J. (2017). "Ieee. Deep pyramidal residual networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6307–6315.

He, J., Zhao, L., Yang, H., Zhang, M., and Li, W. (2019). Hsi-bert: hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geoscience Remote Sens.* 58, 165–178. doi:10.1109/tgrs.2019.2934760

He, K. M., Zhang, X. Y., Ren, S. Q., and Sun, J. (2016b). Ieee. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 770–778. doi:10.1109/CVPR.2016.90

He, L., Li, J., Plaza, A., and Li, Y. (2016a). Discriminative low-rank gabor filtering for spectral–spatial hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 55, 1381–1395. doi:10.1109/tgrs.2016.2623742

He, L., Li, Y., Li, X., and Wu, W. (2014). Spectral–spatial classification of hyperspectral images via spatial translation-invariant wavelet-based sparse representation. *IEEE Trans. Geoscience Remote Sens.* 53, 2696–2712. doi:10.1109/tgrs.2014.2363682

He, M., Li, B., and Chen, H. (2017). "Multi-scale 3d deep convolutional neural network for hyperspectral image classification," in *IEEE international conference on image processing*, 3904–3908.

He, Y., Tu, B., Jiang, P., Liu, B., Li, J., and Plaza, A. (2024). Igroupss-mamba: Interval group spatial-spectral mamba for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–17. doi:10.1109/tgrs.2024.3502055

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7132–7141.

Jiang, Y., Li, Y., Zou, S., Zhang, H., and Bai, Y. (2021). Hyperspectral image classification with spatial consistence using fully convolutional spatial propagation network. *IEEE Trans. Geoscience Remote Sens.* 59, 10425–10437. doi:10.1109/tgrs.2021.3049282

Li, J., Zhao, X., Li, Y., Du, Q., Xi, B., and Hu, J. (2018). Classification of hyperspectral imagery using a new fully convolutional neural network. *IEEE Geoscience Remote Sens. Lett.* 15, 292–296. doi:10.1109/lgrs.2017.2786272

Li, S., Wu, H., Wan, D., and Zhu, J. (2011). An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Syst.* 24, 40–48. doi:10.1016/j.knosys.2010.07.003

Li, X., Ding, M. L., and Pizurica, A. (2020). Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 58, 2615–2629. doi:10.1109/tgrs.2019.2952758

Licciardi, G., Marpu, P. R., Chanussot, J., and Benediktsson, J. A. (2012). Linear versus nonlinear pca for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geoscience Remote Sens. Lett.* 9, 447–451. doi:10.1109/lgrs.2011.2172185

Liu, H., Li, W., Xia, X. G., Zhang, M., Gao, C. Z., and Tao, R. (2022). Central attention network for hyperspectral imagery classification. *IEEE Trans. Neural Netw. Learn. Syst.* 34, 8989–9003. doi:10.1109/tnnls.2022.3155114

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lu, Z., Xu, B., Sun, L., Zhan, T., and Tang, S. (2020). 3-d channel and spatial attention based multiscale spatial–spectral residual network for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 4311–4324. doi:10.1109/jstars.2020.3011992

Ma, W., Yang, Q., Wu, Y., Zhao, W., and Zhang, X. (2019). Double-branch multi-attention mechanism network for hyperspectral image classification. *Remote Sens.* 11, 1307. doi:10.3390/rs11111307

Makantasis, K., Karantzalos, K., Doulamis, A., and Doulamis, N. (2015). Deep supervised learning for hyperspectral data classification through convolutional neural networks. *IEEE Int. Geoscience Remote Sens. Symposium (IGARSS)*, 4959–4962. doi:10.1109/igarss.2015.7326945

Mei, X., Pan, E., Ma, Y., Dai, X., Huang, J., Fan, F., et al. (2019). Spectral-spatial attention networks for hyperspectral image classification. *Remote Sens.* 11, 963. doi:10.3390/rs11080963

Mohanty, P., Panditrao, S., Mahendra, R., Kumar, S., and Kumar, T. S. (2016). Identification of coral reef feature using hyperspectral remote sensing. *Multispectral, Hyperspectral, Ultraspectral Remote Sens. Technol. Tech. Appl. VI* 9880, 311–321. doi:10.1117/12.2227991

Pang, L., Yao, J., Li, K., and Cao, X. (2025). Special: zero-shot hyperspectral image classification with clip. *arXiv Prepr. arXiv:2501*, 16222. doi:10.48550/arXiv.2501.16222

Paoletti, M. E., Haut, J. M., Fernandez-Beltran, R., Plaza, J., Plaza, A. J., and Pla, F. (2018). Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 57, 740–754. doi:10.1109/tgrs.2018.2860125

Paoletti, M. E., Haut, J. M., Fernandez-Beltran, R., Plaza, J., Plaza, A. J., and Pla, F. (2019). Deep pyramidal residual networks for spectral-spatial hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 57, 740–754. doi:10.1109/tgrs.2018.2860125

Pooja, K., Nidamanuri, R. R., and Mishra, D. (2019). Multi-scale dilated residual convolutional neural network for hyperspectral image classification. *Workshop Hyperspectral Imaging Signal Process. Evol. Remote Sens. (WHISPERS)*, 1–5. doi:10.1109/whispers.2019.8921284

Prasad, S., and Bruce, L. M. (2008). Limitations of principal components analysis for hyperspectral target recognition. *IEEE Geoscience Remote Sens. Lett.* 5, 625–629. doi:10.1109/lgrs.2008.2001282

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogrammetry Remote Sens.* 67, 93–104. doi:10.1016/j.isprsjprs.2011.11.002

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *Med. Image Comput. Computer-Assisted Intervention–MICCAI 2015 18th Int. Conf. Munich, Ger. Oct. 5-9, 2015, Proc. Part III* 18, 234–241. doi:10.1007/978-3-319-24574-4_28

Roy, S. K., Deria, A., Shah, C., Haut, J. M., Du, Q., and Plaza, A. (2023). Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 61, 1–15. doi:10.1109/tgrs.2023.3242346

Sun, M., Zhang, J., He, X., and Zhong, Y. (2024). Bidirectional mamba with dual-branch feature extraction for hyperspectral image classification. *Sensors* 24, 6899. doi:10.3390/s24216899

Tiwari, K. C., Arora, M. K., and Singh, D. (2011). An assessment of independent component analysis for detection of military targets from hyperspectral images. *Int. J. Appl. Earth Observation Geoinformation* 13, 730–740. doi:10.1016/j.jag.2011.03.007

Villa, A., Benediktsson, J. A., Chanussot, J., and Jutten, C. (2011). Hyperspectral image classification with independent component discriminant analysis. *IEEE Trans. Geoscience Remote Sens.* 49, 4865–4876. doi:10.1109/tgrs.2011.2153861

Wang, X., Tan, K., Du, P., Pan, C., and Ding, J. (2022). A unified multiscale learning framework for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 60, 1–19. doi:10.1109/tgrs.2022.3147198

Woo, S. H., Park, J., Lee, J. Y., and Kweon, I. S. (2018). Cbam: convolutional block attention module. *Eur. Conf. Comput. Vis.* 11211, 3–19. doi:10.1007/978-3-030-01234-2_1

Wu, S. F., Zhang, J. P., and Zhong, C. X. (2019). Ieee. Multiscale spectral-spatial unified networks for hyperspectral image classification. *IEEE Int. Geoscience Remote Sens. Symposium (IGARSS)*, 2706–2709. doi:10.1109/igarss.2019.8900581

Yang, K., Sun, H., Zou, C., and Lu, X. (2021a). Cross-attention spectral–spatial network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 60, 1–14. doi:10.1109/tgrs.2021.3133582

Yang, L., Zhang, F., Wang, P. S. P., Li, X., and Meng, Z. (2022b). Multi-scale spatial-spectral fusion based on multi-input fusion calculation and coordinate attention for hyperspectral image classification. *Pattern Recognit.* 122, 108348. doi:10.1016/j.patcog.2021.108348

Yang, L., Zhang, R. Y., Li, L., and Xie, X. (2021b). "Simam: a simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*, 11863–11874.

Yang, X., Cao, W., Lu, Y., and Zhou, Y. (2022a). Hyperspectral image transformer classification networks. *IEEE Trans. Geoscience Remote Sens.* 60, 1–15. doi:10.1109/tgrs.2022.3171551

Yao, J., Zhang, B., Li, C., Hong, D., and Chanussot, J. (2023). Extended vision transformer (exvit) for land use and land cover classification: a multimodal deep learning framework. *IEEE Trans. Geoscience Remote Sens.* 61, 1–15. doi:10.1109/tgrs.2023.3284671

Zhan, L., Ye, P., Fan, J., and Chen, T. (2024). U^{2}ConvFormer: marrying and evolving nested U-net and scale-aware transformer for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–14. doi:10.1109/tgrs.2024.3394901

Zhang, J., Meng, Z., Zhao, F., Liu, H., and Chang, Z. (2022). Convolution transformer mixer for hyperspectral image classification. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3208935

Zhao, D., Yu, P., Guo, F., Yang, X., Ma, Y., Wang, C., et al. (2024a). Classification of hyperspectral images of explosive fragments based on spatial–spectral combination. *Sensors* 24, 7131. doi:10.3390/s24227131

Zhao, J., Wang, J., Ruan, C., Dong, Y., and Huang, L. (2024b). Dual-branch spectral-spatial attention network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–18. doi:10.1109/tgrs.2024.3351997

Zheng, X., Sun, H., Lu, X., and Xie, W. (2022). Rotation-invariant attention network for hyperspectral image classification. *IEEE Trans. Image Process.* 31, 4251–4265. doi:10.1109/tip.2022.3177322

Zheng, Z., Zhong, Y., Ma, A., and Zhang, L. (2020). Fpga: fast patch-free global learning framework for fully end-to-end hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 58, 5612–5626. doi:10.1109/tgrs.2020.2967821

Zhong, Z., Li, J., Luo, Z., and Chapman, M. (2017). Spectral–spatial residual network for hyperspectral image classification: a 3-d deep learning framework. *IEEE Trans. Geoscience Remote Sens.* 56, 847–858. doi:10.1109/tgrs.2017.2755542

Zhou, J., Sheng, J., Ye, P., Fan, J., He, T., Wang, B., et al. (2024). Exploring multi-timestep multi-stage diffusion features for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–16. doi:10.1109/tgrs.2024.3407206

Zhu, M. H., Jiao, L. C., Liu, F., Yang, S. Y., and Wang, J. N. (2021b). Residual spectral-spatial attention network for hyperspectral image classification. *IEEE Trans. Geoscience Remote Sens.* 59, 449–462. doi:10.1109/tgrs.2020.2994057

Zhu, Q., Deng, W., Zheng, Z., Zhong, Y., Guan, Q., Lin, W., et al. (2021a). A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification. *IEEE Trans. Cybern.* 52, 11709–11723. doi:10.1109/tcyb.2021.3070577