# Multilevel Feature Gated Fusion Based Spatial and Frequency Domain Attention Network for Joint Classification of Hyperspectral and LiDAR Data

Cuiping Shi, *Member, IEEE*, Zhipeng Zhong, Shihang Ding, Yeqi Lei, Liguo Wang, *Member, IEEE*, and Zhan Jin

***Abstract*—Hyperspectral images provide rich spectral information, while LiDAR data supplements three-dimensional spatial structural information. The combination of the two can effectively improve the accuracy of land cover classification. However, how to effectively utilize their complementary advantages for cross modal feature fusion and enable the fused joint features to capture global contextual information while maintaining local texture details is a challenge. In addition, most existing joint classification methods based on attention and transformer only perform global modeling in the spatial domain, ignoring the sensitivity of the frequency domain to fine features. In this article, a multilevel feature gated fusion based spatial and frequency domain attention network is proposed for joint classification of hyperspectral and LiDAR data. First, extract multilevel convolutional features from hyperspectral and LiDAR images and adaptively fuse them through a gating mechanism. Then, design an attention module that combines spatial frequency domain to model global fine features. In addition, a carefully designed texture feature extraction module is utilized to further enhance local fine feature extraction. The experimental results on three commonly used datasets show that the classification performance of the proposed method is significantly better than some state-of-the-art methods.**

***Index Terms*—Attention, classification, deep learning, hyperspectral imaging (HSI), LiDAR data.**

## I. Introduction

BY COMBINING the advantages of different sensors, modern remote sensing technology makes the classification and recognition of ground objects achieve unprecedented accuracy and efficiency. Among them, hyperspectral images (HSI) provide rich spectral information, while LiDAR supplements detailed information on spatial three-dimensional structures. HSI has dozens to hundreds of bands and can capture the reflection characteristics of surface objects at different wavelengths [1]. This almost continuous spectral resolution enables hyperspectral data to perform well in identifying land features with subtle spectral differences and is widely used in fields such as marine hydrological exploration [2] and precision agriculture [3]. However, the limitation of hyperspectral data is that it mainly provides two-dimensional plane information and lacks height data of land features. Therefore, it is necessary to introduce LiDAR data. LiDAR technology measures distance by actively emitting laser pulses and receiving their echoes, thereby obtaining elevation information of ground objects. LiDAR can not only penetrate clouds and vegetation, but also generate digital elevation models, providing precise measurements in the vertical direction. This feature makes it a powerful tool in fields such as terrain surveying [5], forestry monitoring [6], meteorological observation [7], and urban planning [8]. Combining HSI and LiDAR data can further enhance the accuracy and reliability of land cover classification. Therefore, researching image classification methods based on the fusion of HSI and LiDAR information has gradually become a hot topic.

In the early work of HSI and LiDAR data classification, most methods were based on machine learning to extract shallow features for land cover classification. For example, some methods are based on support vector machines, Gaussian maximum likelihood [9], and random forests [10]. In [11], some extended morphological attribute contours were designed for joint classification. Ghamisi et al. [12] utilized attribute contours to extract spatial information as features for classification. A graph based fusion method [13] has been proposed, which improves the extraction method of morphological contours. Rasti et al. [14] Combined with extinction profile joint feature extraction and total variation analysis. However, the above work relies on manually designed algorithms to select and extract features, overly relying on prior knowledge, and cannot adaptively summarize the intrinsic features of remote sensing data.

In recent years, the rapid development of deep learning has promoted innovation in the field of remote sensing image processing technology [15], [16], [17], [18], [19]. Convolutional neural networks (CNNs) have shown great potential in remote sensing image classification due to their powerful feature extraction and deep semantic automatic learning capabilities [20], [21], [22]. For example, [23] proposed a coupled CNN and

Received 6 November 2024; revised 10 January 2025; accepted 18 January 2025. Date of publication 3 February 2025; date of current version 25 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409, in part by the Science and Technology Plan Project of Huzhou under Grant 2024GZ36, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 145109145. *(Corresponding author: Cuiping Shi.)*

Cuiping Shi, Zhipeng Zhong, Shihang Ding, and Yeqi Lei are with the College of Information Engineering, Huzhou University, Huzhou 313000, China (e-mail: shicuiping@zjhu.edu.cn; 2023388015@stu.zjhu.edu.cn; 2023388002@stu.hzsf.end.cn; 2023388418@stu.zjhu.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China (e-mail: wangliguo@hrbeu.edu.cn).

Zhan Jin is with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: jinzhan@qqhru.edu.cn).

The codelink of the proposed method is https://github.com/Zzp-12/GFSFN.

Digital Object Identifier 10.1109/JSTARS.2025.3534286

weighted summation strategy to improve joint classification accuracy, while [24] employed a three-branch neural network to enhance and learn spatial and spectral features. In response to the limitation of traditional CNN convolution kernels having a single scale, [25] proposed a classification model based on dilated convolution, which demonstrates the advantages of dilated convolution in hyperspectral image processing. To alleviate the possible overfitting caused by deep CNNs, [26] proposed an activation function called Lush and established a multilayer feature fusion bias network based on it. To enhance the interpretability of remote sensing data fusion, [27], [28] explores feature fusion from the perspective of geometric structure. However, although CNN based methods perform well in extracting local spatial features, they cannot fully utilize the global sequence characteristics of spectral features.

Transformer has become an important tool in fields such as natural language processing [29] and computer vision [30] due to its powerful global modeling capabilities. Thanks to its unique attention mechanism, transformer can effectively improve the long distance feature extraction ability and recognition performance of the model. Therefore, transformer was introduced into the field of remote sensing image processing, and some models that combine CNN and attention mechanisms [31], [32], [33], [34] were proposed, demonstrating excellent classification performance. For example, [32] proposed a feature complementary attention network based on adaptive knowledge filtering to address the issue of redundant information in HSI feature extraction. Ding et al. [33] proposes a global local transformer that learns and fuses spatial and spectral features from multiple input scales. To improve the ability to extract key information features, [34] proposed a dual multiscale adaptive attention mechanism. Considering the complementarity of multisource data, [35] proposes a hierarchical mutual-assistance learning mechanism based on height information to enhance modality specific features. In the task of joint classification, some attention mechanisms have also been widely studied. For example, methods such as [36] were based on squeeze-and-excitation (SE) attention to adaptively adjust the information exchange of modal features, while some methods have improved traditional attention. Specifically, [37] uses graph convolutional networks to construct attention weights, while [38] directly takes features from different modalities as input and participates in the calculation process of attention weights. However, SE attention focuses on modeling the relationships between channels and lacks effective attention to spatial information. Although traditional attention can capture global dependencies, its matrix operation method can lead to high computational complexity.

To further improve the computational efficiency and generalization ability of the model while preserving important frequency domain information of features, some methods introduce fast Fourier transform (FFT) [39], [40]. FFT has the characteristics of parameter free and fast computation, which can help models effectively learn features in the frequency domain. Combining it with deep learning can improve the model's ability to process complex data. Most existing joint classification methods based on attention mechanisms mainly focus on learning global features in the spatial domain, often neglecting the importance of frequency domain features. In fact, the features extracted from the frequency domain can comprehensively cover all frequency ranges, thus better capturing short-range and long-range dependencies. More importantly, due to its sensitivity to subtle changes, the frequency domain can identify those small but crucial frequency variations of the signal, providing richer and more detailed information. Therefore, exploring how to effectively utilize frequency domain features into some network architectures such as CNN and attention may become a key to further improving joint classification performance.

Based on the above analysis, this article proposes a multilevel feature gated fusion based spatial and frequency domain attention network (GFSFN) for joint classification of hyperspectral and LiDAR data. First, extract single modal multilevel dilated convolution features and fuse spatial features, and then perform adaptive fusion through a gating mechanism module. Second, design a spatial frequency domain attention module that maps features to the frequency domain through FFT to obtain attention weights, and combines the original spatial domain features to effectively model global fine features. In addition, the texture feature extraction module based on convolution has a significant effect in capturing local fine features.

The main contributions of this article are as follows.

1) To fully integrate the features of hyperspectral and LiDAR images, a two-stage multilevel cross modal feature gating fusion (CMGF) module was carefully designed. In the first stage, through the cross modal space feature (CMSF) module, the dilated convolution features of two modalities at the same channel depth were fused to reduce the differences between heterogeneous data and achieve effective cross modal fusion. In the second stage, a multilevel feature gated fusion (MLGF) module was constructed. This module can finely control the joint features of different depths obtained in the previous stage, and adaptively adjust the proportion of feature flow from each layer to the next stage. This design makes the feature fusion process more flexible and efficient, thereby further improving the performance and accuracy of the model.
2) A spatial and frequency domain attention (S&FA) module was constructed to accurately model global features and long-range dependencies. Specifically, construct an attention weight matrix in the frequency domain and map it back to the spatial domain to combine it with the original detail information. By learning the interaction between frequency and spatial domain features, this model can capture the global structure of joint features, especially fine features, thereby improving the richness and accuracy of feature representation.
3) A convolutional based texture feature extraction module (CTEM) is proposed to effectively capture local texture features, thereby further enhancing the network's ability to extract fine local features and providing more detailed and robust feature support for classification.

## II. Methodology

The joint classification framework of HSI and LiDAR data proposed in this article is shown in Fig. 1, which mainly consists of three stages. First, two single-model features are extracted and

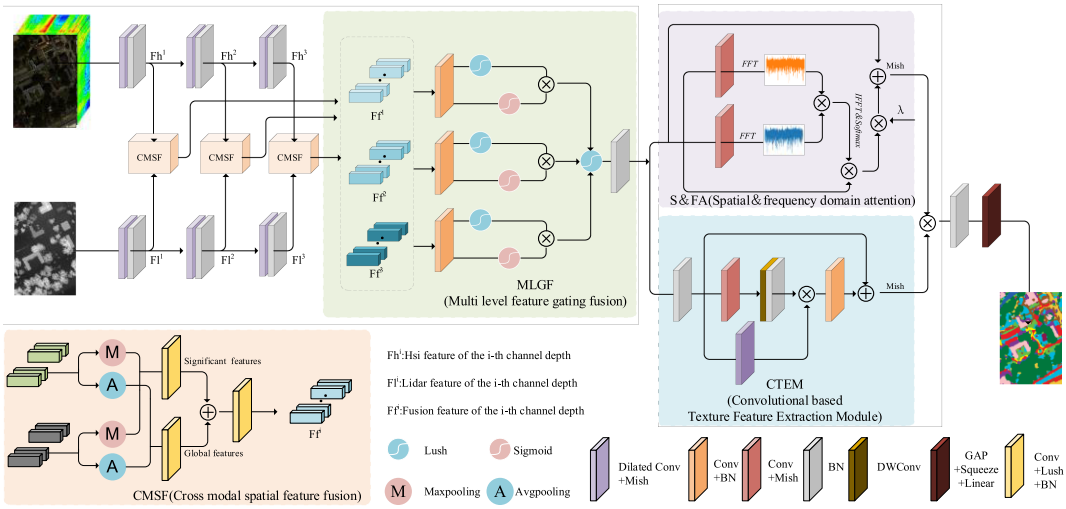



Fig. 1. Overall structure of the proposed GFSFN.

fused. Then, fine features are modeled through global local dual branches. Finally, the features are input into the classification head. Let $Xh \in R^{H\times W\times 30}$ and $Xl \in R^{H\times W\times C}$ be the patches for HSI and LiDAR, respectively, where $H\times W$ is the spatial size and $C$ represents the number of channels in the LiDAR patch. First, perform PCA dimensionality reduction on the HSI patch. To obtain more receptive fields, three dilated convolution blocks are utilized to extract multilevel convolution features from the input HSI patch and LiDAR patch, with channel depths of 32, 64, and 128, respectively. The calculation process can be described as follows:

$$Fm^1 = \mathrm{BN}\left(\mathrm{Mish}\left(\mathrm{DConv}\left(Xm\right)\right)\right) \quad (1)$$

$$Fm^{\mathrm{i}+1} = \mathrm{BN}\left(\mathrm{Mish}\left(\mathrm{DConv}\left(Fm^{\mathrm{i}}\right)\right)\right). \quad (2)$$

Among them, BN(·), Mish(·), and DConv(·) represent batch normalization, activation function Mish, and dilated convolution, respectively. $i\in\{1,2\}$, $m\in\{h,l\}$, $h$ and $l$ represent HSI and LiDAR, and $Fm^i$ represents the characteristics of mode $m$ at the depth of the $i$th layer channel. A detailed introduction to the details of each module is provided as follows.

### A. CMGF

To fully integrate the multilevel convolutional features extracted from HSI and LiDAR, a CMGF module was designed, which is a two-stage feature fusion process. To reduce the differences in heterogeneous multimodal data, considering the semantic differences contained in different depth features, the spatial features of two modalities are first fused at the same channel depth. The calculation process of CMSF can be represented as follows:

$$Fg^i = BN(\mathrm{Lush}\left(\mathrm{Conv}\left(\mathrm{Avg}\left(Fh^i\right)+\mathrm{Avg}\left(Fl^i\right)\right)\right) \quad (3)$$

$$Fs^i = \mathrm{BN}(\mathrm{Lush}\left(\mathrm{Conv}\left(\mathrm{Max}\left(Fh^i\right)+\mathrm{Max}\left(Fl^i\right)\right)\right) \quad (4)$$

$$Ff^i = \mathrm{BN}\left(\mathrm{Lush}\left(\mathrm{Conv}\left(Fg^i+Fs^i\right)\right)\right). \quad (5)$$

Among them, Avg(·), Max(·), Lush(·), and Conv(·) represent average pooling, max pooling, Lush activation function, and standard convolution, respectively. To reduce modal differences, a cross-modal fusion of spatial features is performed. $Fg^i$ represents the overall spatial features of the $i$th layer, which is obtained by performing average pooling operations on the spectral dimension 30 of HSI features and the channel dimension C of LiDAR features through convolutional blocks. Similarly, $Fs^i$ represents the significant spatial features of the $i$th layer obtained by max pooling. To enhance the representation ability of features, the overall spatial features and salient spatial features are summed to extract convolutional features, resulting in three joint features $Ff^i$ with different channel depths. However, simple addition operations cannot effectively fuse these joint features at different depths, thus failing to fully utilize the different semantic features at different depths. Inspired by the gating concept in long short-term memory networks, in this article, an MLGF is proposed to adaptively fuse joint features of different depths. The calculation process of MLGF can be described as follows:

$$\begin{aligned} Ff^{i'} &= \mathrm{Lush}(\mathrm{BN}\left(\mathrm{Conv}\left(Ff^i\right)\right) \\ &\quad \times \mathrm{Sigmoid}\left(\mathrm{BN}\left(\mathrm{Conv}\left(Ff^i\right)\right)\right) \\ Ff'' &= \mathrm{BN}\left(\mathrm{Lush}\left(Ff^{1'}+Ff^{2'}+Ff^{3'}\right)\right). \end{aligned} \quad (6)$$

Among them, Sigmoid(·) represents the activation function Sigmoid. The joint features $Ff^i$ at all levels are unified in terms of channel number through convolution and batch normalization to facilitate further fusion. In addition, in our previous work [26], we proposed a Lush function that can effectively prevent network overfitting and has good generalization performance. This article uses Lush as the activation function to increase nonlinear expression. Sigmoid is adopted as the weight generator to generate weights that control the importance of features at this level. Following, nonlinear features multiply with weights to obtain the final output features $Ff^{i\prime}$ at this level. The three levels of features are added together, and after Lush and batch

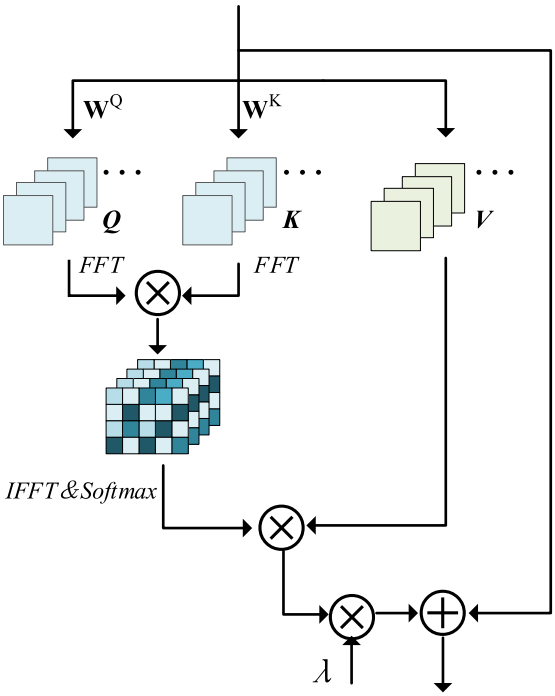



Fig. 2. Structure diagram of S&FA.

normalization, the joint feature $Ff''$ is obtained and sent to the next layer.

### B. S&FA

The attention mechanism simulates the way the human visual system processes information, allowing the model to focus more on key parts of the input information and effectively improve the classification performance of the network. However, most existing attention mechanisms for joint classification methods only model globally in the spatial domain, while features in frequency domain also have significant value and potential that cannot be ignored. Considering the complexity of matrix multiplication in traditional attention, this paper proposes a S&FA module, whose structure is shown in Fig. 2. The process of S&FA can be represented as follows:

$$Q = \text{Mish}\left(\text{Conv}\left(Ff''\right)\right) \tag{7}$$

$$K = \text{Mish}\left(\text{Conv}\left(Ff''\right)\right) \tag{8}$$

$$Attn = \text{FFT}\left(Q\right) \times \text{FFT}\left(K\right) \tag{9}$$

$$\text{Out}A = \text{Softmax}\left(\text{IFFT}\left(\text{Att}\right)\right) \times V \tag{10}$$

$$FG = \lambda \times \text{Out}A + Ff'' . \tag{11}$$

Among them, Softmax(·), FFT(·), and IFFT(·) represent the activation function Softmax, FFT, and inverse fast Fourier transform (IFFT), respectively. The original joint feature $Ff''$ is used as the module input, and it is convolved to obtain *Q* and *K*, which are mapped to the frequency domain through FFT and multiplied element by element to obtain the weight matrix *Attn*. Map *Attn* back to the spatial domain through IFFT and multiply it with the original feature $Ff''$ as *V* to obtain the output Out*A*. In addition,

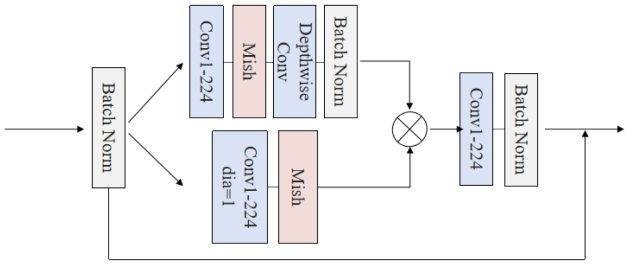



Fig. 3. Structure diagram of CTEM.

λ is adopted as an automatically learned parameter in the model to control the strength of attention features, and skip connections are utilized in the final stage to prevent network overfitting. *FG* is the final output of the S&FA module.

### C. CTEM

Local details are crucial for understanding image categories, as capturing subtle local features often determines the model's ability to accurately identify and distinguish similar categories. This article proposes a CTEM module, which structure is shown in Fig. 3. The process of CTEM can be described as follows:

$$Fdw = \text{DwConv}\left(\text{Mish}\left(\text{Conv}\left(\text{BN}\left(Ff''\right)\right)\right)\right) \tag{12}$$

$$Fd = \text{Mish}\left(\text{DConv}\left(\text{BN}\left(Ff''\right)\right)\right) \tag{13}$$

$$FL = \text{BN}\left(\text{Conv}\left(Fdw \times Fd\right)\right) + \text{BN}\left(Ff''\right). \tag{14}$$

DwConv (·) represents depth-wise convolution. The multiplication of two sets of deep features, *Fdw* and *Fd*, can effectively enhance the weight of important local features in the joint features, thereby significantly improving classification accuracy. Finally, residual branches are used to avoid network degradation, and *FL* is the final output of the CTEM module.

### D. Classification Head

To improve the nonlinear expression ability of the network, the global fine feature *FG* and the local fine feature *FL* are multiplied after Mish activation, and then class prediction is performed. The calculation process can be described as follows:

$$Ycls = FC(\text{GAP}\left(\text{Mish}\left(FL\right) \times \text{Mish}\left(FG\right)\right). \tag{15}$$

Among them, *FC* (·) and GAP (·) represent linear layer and global average pooling, respectively, while *Ycls* is the category predicted for the current input. Training with cross entropy loss, the calculation process of the loss *Lcls* can be described as follows:

$$Lcls = -\sum Yt \cdot \log\left(Ycls\right) \tag{16}$$

where *Yt* is the ground-truth label.

## III. Experimental Results and Analysis

To evaluate the effectiveness of the proposed GFSFN method, a lot of experiments were performed on three commonly used datasets. Some experimental comparisons between the proposed

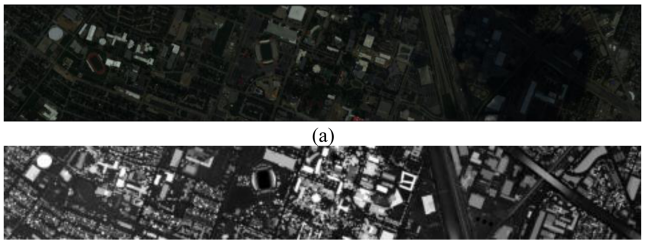


Fig. 4. Houston2013 dataset. (a) Pseudo color image of HSI. (b) LiDAR image.

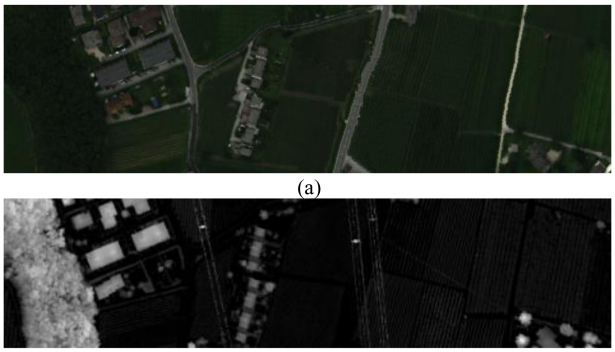


Fig. 5. Trento dataset. (a) Pseudo color image of HSI. (b) LiDAR image.

method and other advanced methods were conducted. The experimental results demonstrate the effectiveness of the proposed method, and the datasets used in the experiment include Houston dataset, Trento dataset, and MUUFL dataset.

### A. Dataset Description

*1) Houston 2013 Dataset:* This dataset was obtained by the National Center for Airborne Laser Mapping in June 2012 on the University of Houston campus and adjacent urban areas, and was provided by the IEEE GRSS Data Fusion Competition. HSI includes 144 bands with a wavelength range of 0.38–1.05 $\mu$m. And LiDAR data are represented by a single band. The spatial size of the HSI and LiDAR datasets is 349 $\times$ 1905 pixels, with a spatial resolution of 2.5 m. This dataset contains 15 different categories, totaling 15 029 real samples. Fig. 4 shows the pseudo color image of HSI and the grayscale image of LiDAR.

*2) Trento Dataset:* This dataset was collected in a rural area located in the southern part of Trento, Italy. HSI has 63 spectral bands with a wavelength range from 0.42 to 0.99 $\mu$m. LiDAR data are represented by a single band. The spatial size of the HSI and LiDAR datasets is 166 $\times$ 600 pixels, with a spatial resolution of 1 m. This dataset contains 6 different categories, totaling 30 214 real samples. Fig. 5 shows the pseudo color image of HSI and the grayscale image of LiDAR.

*3) MUUFL Dataset:* This dataset was obtained in November 2010 in the campus area of South Mississippi Bay Park University. HSI has 72 spectral bands with a wavelength range from 0.38 to 1.05 $\mu$m. LiDAR data are represented by two bands with a wavelength of 1.06 $\mu$m. The spatial size of the HSI and LiDAR datasets is 325 $\times$ 220 pixels. This dataset contains 11 different categories, totaling 53467 real samples. Fig. 6 shows the pseudo color image of HSI and the grayscale image of LiDAR.

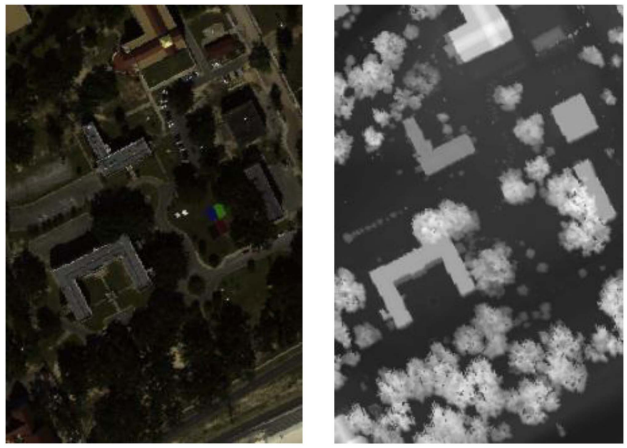


Fig. 6. MUUFL dataset. (a) Pseudo color image of HSI. (b) LiDAR image.

### B. Experimental Setup

The length and width of HSI and LiDAR patches are set to 11, and the number of channels in HSI patch is reduced to 30 after PCA. The LiDAR Patch has one channel for the Houston and Trento datasets, and two channels for the MUUFL dataset. The experiments of this method and other deep learning methods were implemented on the PyTorch platform, with a CPU of i5-12400F, a GPU of NVIDIA RTX 4060Ti, and 32 GB of RAM. To optimize the network, Adam optimizer was chosen as the initial optimizer with a learning rate of 0.0001. For the training phase, the batch size and training epochs are set to 64 and 100, respectively.

This article uses four common evaluation metrics to evaluate the classification performance of the model, namely overall accuracy (OA), average accuracy (AA), kappa coefficient (KAPPA), and accuracy per class. For each indicator, higher values indicate more accurate classification.

### C. Comparison of Classification Performance

To validate the effectiveness of the proposed GFSFN model, we compared it with several state-of-the-art joint classification models for HSI and LiDAR, including FusAtNet [41], $S^2$ENet [36], GLT [33], MFT [42], HCT [43], CALC [44], GAMF [45], and CrossHL [46]. According to the parameter settings described in the original paper, the experimental results are the average of ten independent runs of the network.

Tables I–III provide the classification accuracy of each method. The best results are highlighted in bold. On the Houston 13 dataset, our model achieved the best performance in 10 out of 15 categories, particularly in categories such as residential, commercial, and railway. This achievement is attributed to the S&FA in our network. This module enhances the model's ability to compare and learn complex data by learning the interaction between spectral and spatial features, thereby significantly

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON THE HOUSTON 2013 DATASET

| No. | Class(Train/Test) | performance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FusAtNet | $S^2$ENet | GLT | MFT | HCT | CALC | GAMF | CrossHL | Ours |
| 1 | Healthy grass (20/1231) | 93.43 | 86.17 | 87.24 | 93.31 | 96.60 | 90.01 | 94.05 | **97.51** | 94.50 |
| 2 | Stressed grass (20/1234) | 89.81 | 98.60 | 99.02 | 94.94 | 97.05 | 97.97 | 96.99 | 96.18 | **99.79** |
| 3 | Synthetis grass (20/677) | 99.70 | 99.63 | **100** | 99.05 | 99.38 | 99.11 | 99.48 | 99.56 | 99.79 |
| 4 | Tree (20/1224) | 95.74 | 93.40 | 97.79 | 92.78 | 97.60 | 92.48 | 95.02 | 93.82 | **99.30** |
| 5 | Soil (20/1222) | 98.67 | 99.85 | 99.83 | 98.48 | **100** | **100** | 99.27 | **100** | **100** |
| 6 | Water (20/305) | 94.59 | 97.57 | 94.42 | 96.56 | 97.64 | 95.74 | 97.44 | 95.51 | **98.46** |
| 7 | Residential (20/1248) | 88.88 | 93.97 | 87.25 | 96.16 | 88.89 | 97.12 | 95.12 | 93.75 | **97.95** |
| 8 | Commercial (20/1224) | 83.34 | 74.79 | 92.56 | 77.07 | 86.31 | 82.03 | 71.92 | 76.54 | **95.71** |
| 9 | Road (20/1232) | 54.38 | 83.74 | **96.99** | 77.43 | 81.49 | 84.66 | 79.81 | 79.81 | 92.14 |
| 10 | Highway (20/1207) | 72.39 | 90.82 | 98.01 | 94.72 | 97.27 | **99.75** | 93.10 | 93.05 | 98.38 |
| 11 | Railway (20/1215) | 83.54 | 92.58 | 96.04 | 85.28 | 96.86 | 96.54 | 85.79 | 89.61 | **97.19** |
| 12 | Park lot 1 (20/1213) | 75.49 | 91.81 | 92.41 | 91.58 | 92.09 | 89.37 | 92.06 | **93.20** | 92.52 |
| 13 | Park lot 2 (20/449) | 87.42 | 92.65 | 98.88 | 98.08 | 96.79 | 96.44 | 93.85 | 96.97 | **99.29** |
| 14 | Tennis court (20/408) | 92.21 | **100** | 99.97 | 99.56 | 99.88 | **100** | 99.56 | 99.95 | **100** |
| 15 | Running track (20/640) | 90.34 | 99.98 | **100** | 99.77 | 99.95 | 99.53 | **100** | 99.69 | **100** |
| | OA (%) | 85.21 | 91.87 | 95.44 | 91.64 | 94.33 | 93.91 | 91.67 | 92.58 | **97.23** |
| | AA (%) | 86.66 | 93.04 | 96.02 | 92.98 | 95.19 | 94.72 | 92.90 | 93.68 | **97.67** |
| | KAPPA*100 | 84.00 | 91.21 | 95.07 | 90.96 | 93.87 | 93.41 | 91.00 | 91.98 | **97.01** |

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON THE TRENTO DATASET

| No. | Class(Train/Test) | performance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FusAtNet | $S^2$ENet | GLT | MFT | HCT | CALC | GAMF | CrossHL | Ours |
| 1 | Apple trees (20/4014) | 96.57 | 98.82 | 98.97 | 98.29 | 98.94 | 99.57 | 98.85 | 99.27 | **99.86** |
| 2 | Buildings (20/2883) | **99.28** | 96.51 | 98.30 | 95.06 | 96.53 | 98.67 | 98.37 | 97.43 | 98.48 |
| 3 | Ground (20/459) | 97.84 | 99.78 | 99.37 | 98.37 | 97.67 | 98.61 | **100** | 99.83 | 98.00 |
| 4 | Woods (20/9103) | 99.92 | 99.84 | **100** | 99.98 | **100** | **100** | 99.88 | 99.99 | **100** |
| 5 | Vineyard (20/10481) | 98.18 | 99.94 | **100** | 99.95 | 99.15 | **100** | 99.95 | 99.94 | 99.98 |
| 6 | Roads (20/3154) | 91.43 | 93.24 | 95.13 | 92.29 | 95.31 | 94.37 | 93.50 | 88.48 | **98.05** |
| | OA (%) | 97.88 | 98.72 | 99.15 | 98.44 | 98.70 | 99.20 | 98.96 | 98.42 | **99.59** |
| | AA (%) | 97.20 | 98.02 | 98.39 | 97.32 | 97.93 | 98.54 | 98.43 | 97.49 | **99.06** |
| | KAPPA*100 | 97.19 | 98.30 | 98.41 | 97.92 | 98.27 | 98.94 | 98.61 | 97.90 | **99.46** |

improving classification accuracy. On the Trento dataset, our model achieved high accuracy of 99.86% on apple trees and 100% on tree categories, respectively. This outstanding performance is attributed to the CTEM module we designed, which has a high sensitivity to complex textures. Subsequently, the effectiveness of the module will be further validated and demonstrated through visualization analysis of the heat map. The MUUFL dataset is the most challenging among the three datasets, nonetheless, our proposed method still achieves the best results. This is due to our multilevel gating fusion mechanism, which effectively reduces the differences between heterogeneous data and promotes cross modal fusion.

Overall, on the three datasets, the classification accuracy of the proposed method is significantly higher than other advanced methods, and the OA of the proposed method is 1.79%, 0.39%, and 0.53% higher than those of the suboptimal method, respectively, demonstrating the superiority of the proposed method. Among these comparison methods, FusAtNet based on attention mechanism and graph attention performs poorly in GAMF. In transformer based methods, the classification performance of GLT is significantly better than that of MFT and HCT due to the input of multiple spatial scales. However, in the MUUFL dataset, the OA of CrossHL is higher than that of GLT, which proves the effectiveness of the cross modal attention mechanism. Overall, models based on the local-global approach, such as GLT, HCT, and the model proposed in this article, have shown relatively good classification performance. Figs. 7–9 show the classification maps of each method. It can be seen that compared with other methods, the classification performance of our method is better, retaining more texture details, especially in the complex area on the right side of Houston 2013 that is obscured by clouds and mist. The quantitative results shown in the table also confirm this.

### D. Analysis of Module

Table IV shows the impact of different modules on classification performance. In Case 1, the HSI and LiDAR inputs in the

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON THE MUUFL DATASET

| No. | Class(Train/Test) | performance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FusAtNet | $S^2$ENet | GLT | MFT | HCT | CALC | GAMF | CrossHL | Ours |
| 1 | Trees(20/23226) | 84.52 | 88.00 | 89.23 | 88.46 | 84.04 | 90.89 | 88.32 | **91.68** | 90.17 |
| 2 | Mostly grass(20/4250) | 68.96 | 75.15 | **79.07** | 70.90 | 75.49 | 77.18 | 79.18 | 67.60 | 77.94 |
| 3 | Mixed ground surface(20/6862) | 52.46 | 56.05 | 63.27 | 65.85 | 64.42 | 65.16 | 60.45 | 71.55 | **71.96** |
| 4 | Dirt and sand(20/1806) | 70.23 | 87.65 | 94.32 | 93.47 | 91.85 | **97.95** | 94.13 | 89.39 | 91.91 |
| 5 | Road(20/6667) | 78.20 | 81.76 | 83.67 | 76.42 | 73.80 | 70.24 | 79.20 | 83.70 | **84.03** |
| 6 | Water(20/446) | 96.21 | 99.78 | 99.89 | 99.78 | 99.82 | **100** | 99.78 | 96.23 | 99.93 |
| 7 | Building shadow(20/2213) | 74.02 | 90.19 | **93.55** | 84.36 | 83.43 | 78.22 | 80.98 | 89.77 | 88.59 |
| 8 | Building(20/6220) | 81.46 | 91.45 | 93.50 | 91.52 | 91.28 | 92.17 | 92.54 | 94.23 | **94.35** |
| 9 | Sidewalk(20/1365) | 60.63 | **61.68** | 58.41 | 52.20 | 51.63 | 41.25 | 32.67 | 56.26 | 61.34 |
| 10 | Yellow curb(20/163) | 66.01 | 74.23 | 73.93 | 70.06 | 84.72 | 65.03 | 85.28 | 73.19 | **87.98** |
| 11 | Cloth panels(20/249) | 93.82 | 96.39 | 99.08 | **99.56** | 98.31 | 96.39 | 96.39 | 97.83 | 99.48 |
| | OA（%） | 76.58 | 82.00 | 84.55 | 82.18 | 80.02 | 82.54 | 81.96 | 85.43 | **85.96** |
| | AA（%） | 75.14 | 82.03 | 84.36 | 81.14 | 81.71 | 77.38 | 80.81 | 82.86 | **86.15** |
| | KAPPA*100 | 70.06 | 76.95 | 80.12 | 77.13 | 74.54 | 79.50 | 76.84 | 81.07 | **81.81** |

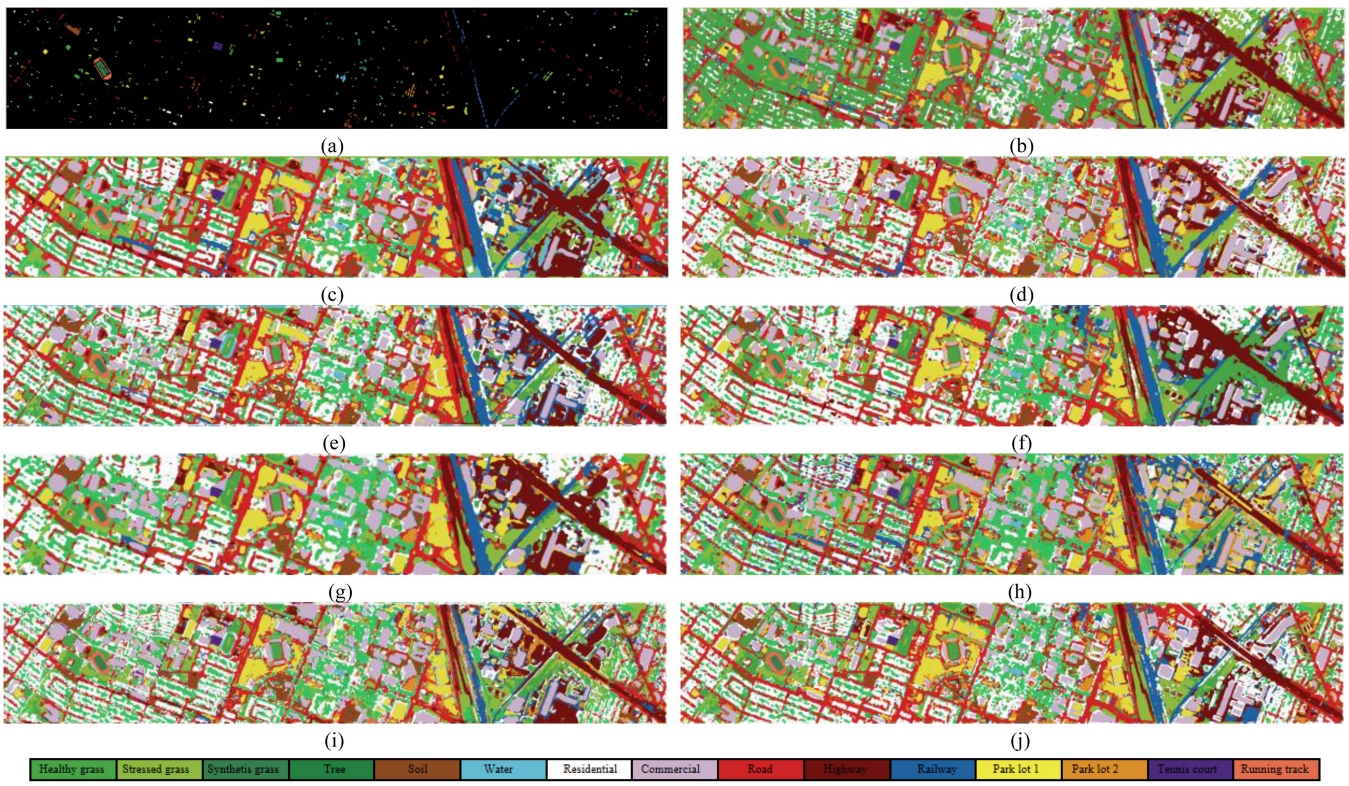



Fig. 7. Classification maps of different methods on the Houston 2013 dataset. (a) Ground true. (b) FusAtNet (85.21%). (c) $S^2$ENet (91.87%). (d) GLT (95.44%). (e) MFT (91.64%). (f) HCT (94.33%). (g) CALC (93.91%). (h) GAMF (91.67%). (i) CrossHL (92.58%). (j) GFSFN (97.23%).

third layer of single-modal features obtained after three dilated convolutional blocks, that is, fh3 and fl3, are directly added into the classification head. Case 2 adds a CMSF module to fuse spatial features of each single modal feature layer, concatenate them according to channel dimensions, and enter the classification head. Thanks to the effectiveness of hierarchical fusion of multilevel dilated convolution features, the addition of CMSF significantly improved the OA on three datasets, especially in the MUFFL dataset where the OA increased by 1.41%. Case 3 adds an MLGF module based on Case 2 to fuse multilevel features instead of concatenation. Add MLGF module to control the proportion of joint feature flow at different depths for adaptive

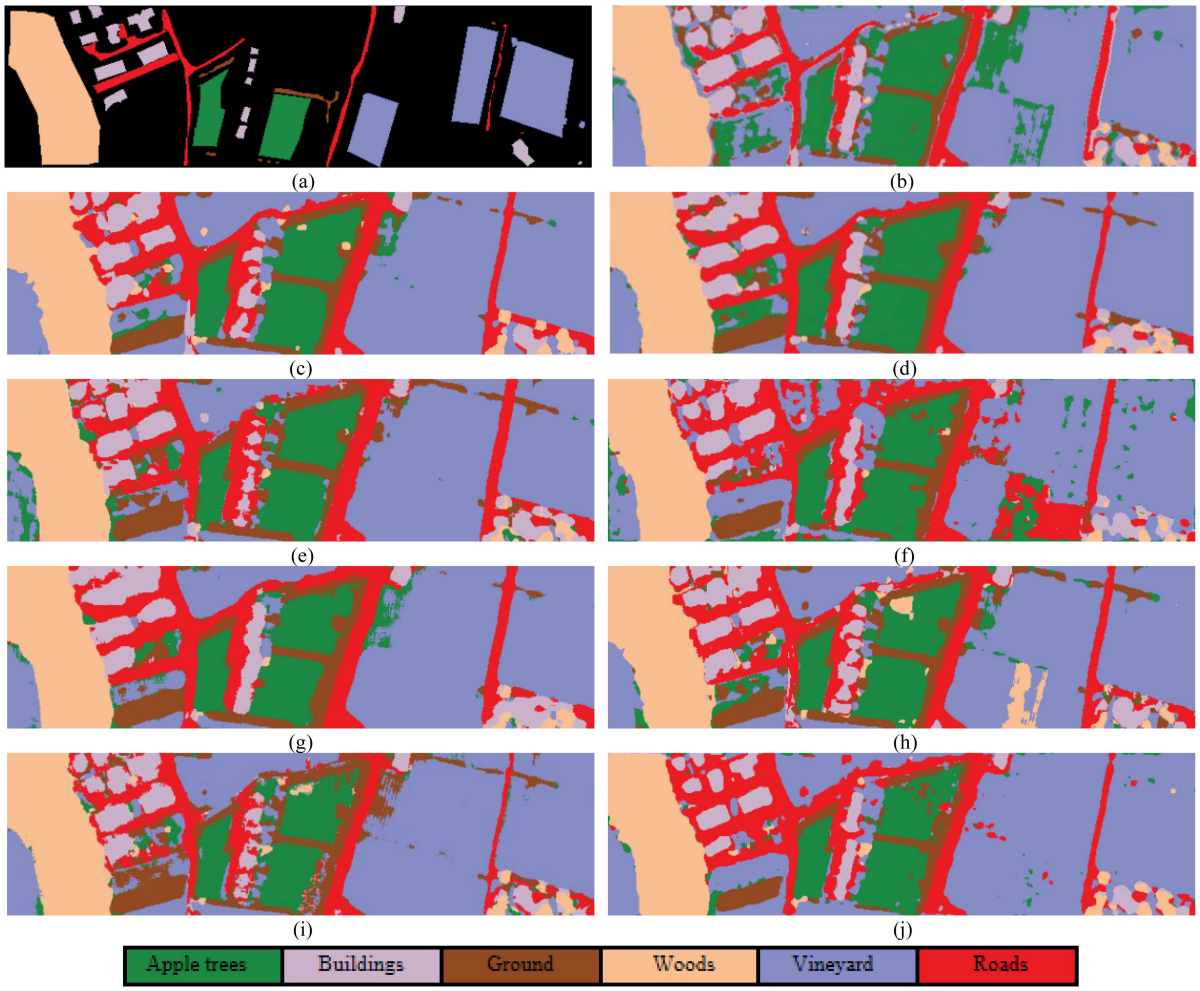



Fig. 8. Classification maps of different methods on the Trento dataset. (a) Ground true. (b) FusAtNet (97.88%). (c) $S^2$ENet (98.72%). (d) GLT (99.15%). (e) MFT (98.44%). (f) HCT (98.70%). (g) CALC (99.20%). (h) GAMF (98.96%). (i) CrossHL (98.42%). (j) GFSFN (99.59%).

TABLE IV
SOME ABLATION EXPERIMENT RESULTS

| Case | component | | | | Houston2013 | | | Trento | | | MUUFL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CMSF | MLGF | CTEM | S&FA | OA | AA | KAPPA | OA | AA | KAPPA | OA | AA | KAPPA |
| 1 | | | | | 94.81 | 95.47 | 94.38 | 98.96 | 98.32 | 98.61 | 83.25 | 83.76 | 78.46 |
| 2 | √ | | | | 96.20 | 96.68 | 95.89 | 99.20 | 98.78 | 98.93 | 84.66 | 85.58 | 80.18 |
| 3 | √ | √ | | | 96.69 | 97.21 | 96.42 | 99.34 | 98.81 | 99.12 | 85.08 | 85.59 | 80.68 |
| 4 | √ | √ | √ | | 96.87 | 97.34 | 96.61 | 99.40 | 98.95 | 99.19 | 85.43 | 85.92 | 81.17 |
| 5 | √ | √ | | √ | 97.02 | 97.47 | 96.78 | 99.54 | 99.03 | 99.39 | 85.33 | 85.70 | 80.97 |
| 6 | √ | √ | √ | √ | **97.23** | **97.67** | **97.01** | **99.59** | **99.06** | **99.46** | **85.96** | **86.15** | **81.81** |

fusion, further improving classification performance. The OA improvement on the three datasets is 0.49%, 0.14%, and 0.42%, respectively. Case 4-5 adds CTEM and S&FA respectively, based on Case 3 to achieve fine modeling of local and global areas, and both modules have significantly improved classification performance. Especially, due to the MUFFL dataset having more dense trees, the CTEM module sensitive to local complex texture features has a greater improvement on OA than S&FA. In the other two datasets, S&FA showed a greater improvement in OA than CTEM. Case 6 has a complete model structure, and compared to Case 1, the OA on three datasets has improved by 2.42%, 0.63%, and 2.71%. Obviously, each module

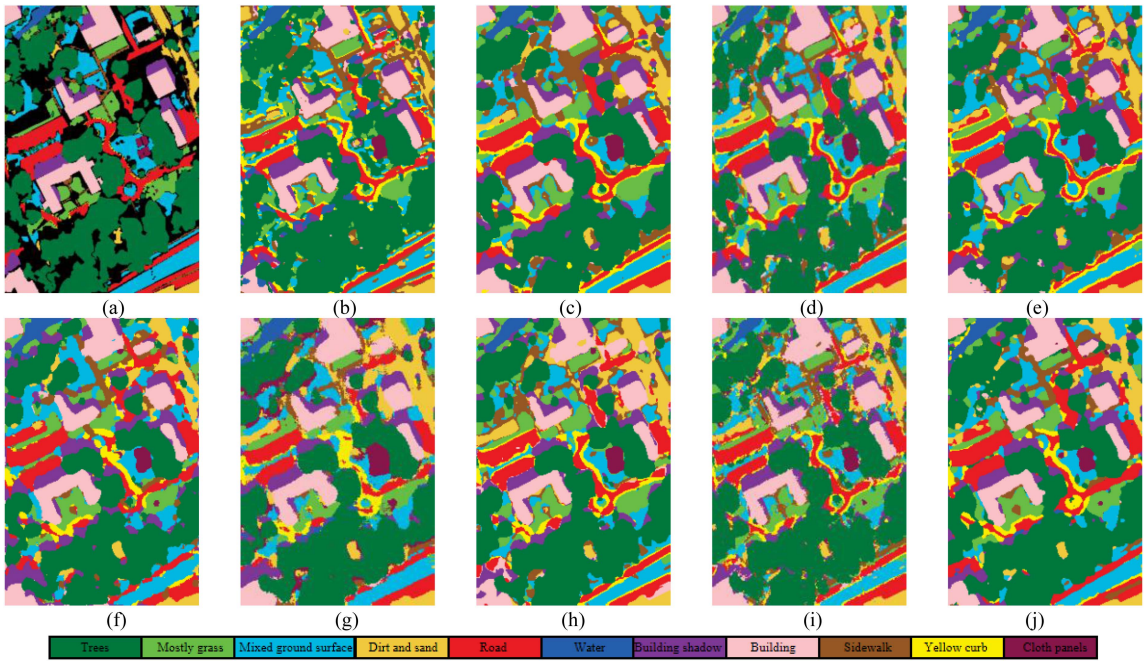



Fig. 9. Classification maps of different methods on the MUUFL dataset. (a) Ground true. (b) FusAtNet (76.58%). (c) $S^2$ENet (82.00%). (d) GLT (84.55%). (e) MFT (82.18%). (f) HCT (80.02%). (g) CALC (82.54%). (h) GAMF (81.96%). (i) CrossHL (85.43%). (j) GFSFN (85.96%).

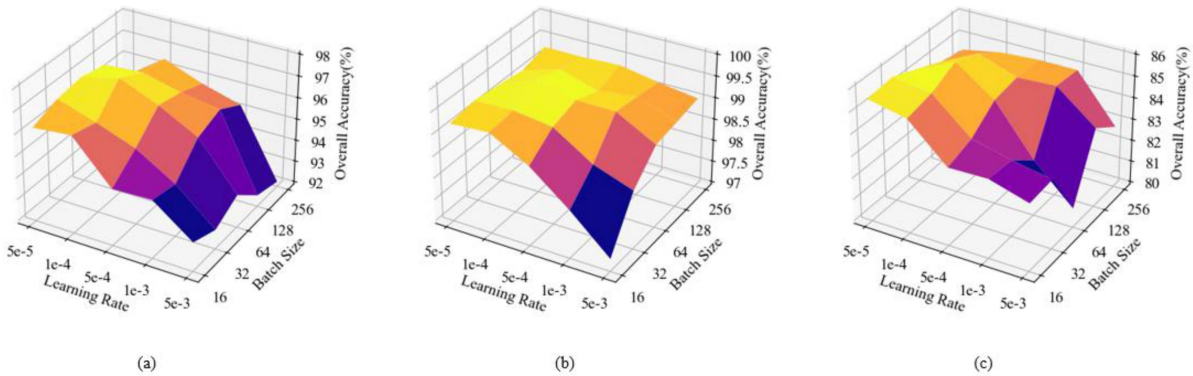



Fig. 10. Impact of learning rate and batch size on OA. (a) Houston 2013. (b) Trento. (c) MUUFL.

has significantly improved classification accuracy, and different modules can effectively work together.

### E. Parameter Analysis

*1) Effect of Different Hyperparameters:* During the process of model training, different combinations of learning rate and batch size can have a significant impact on the classification performance of the model. To select the most suitable learning rate and batch size for GFSFN, some detailed experiments have been conducted on three datasets. Specifically, select the learning rate from {'5e-5', '1e-4', '5e-4', '1e-3', '5e-3'}, and choose the batch size from {'16', '32', '64', '128', '256'}. The experimental results are shown in Fig. 10, which illustrates the changes in model performance under different parameter settings. Fig. 10 shows the experimental results on the Houston 2013, Trento, and MUUFL datasets. In Fig. 10, different colors represent different OA ranges, with yellow indicating the highest OA value and blue indicating the lowest OA value. By observing Fig. 10, it is evident that the model is highly sensitive to different learning rates and batch sizes on the same dataset. This means that selecting appropriate hyperparameters is crucial for improving model performance.

On the Houston 2013 dataset, we observed a significant decrease in model performance as the learning rate increased. On the Trento dataset, when the learning rate reaches 5e-4 and the batch size drops to 32, there is a significant fluctuation in model performance. The MUUFL dataset also shows a similar trend. Overall, the combination of learning rate and batch size within the range of {'5e-5','1e-4'} and {'64 ','128'} performs well, especially when the learning rate is 1e-4 and the batch size is 64, the model performance reaches its optimal level.

TABLE V
COMPARISON OF FLOPS, PARAMETERS, AND TRAINING TIME ON THE HOUSTON 2013 DATASET

| Complexity | FusAtNet | $S^2$ENet | GLT | MFT | HCT | CALC | GAMF | CrossHL | Ours |
|---|---|---|---|---|---|---|---|---|---|
| FLOPs(M) | 3462.66 | 27.08 | 156.54 | 21.06 | 7.93 | 28.75 | 2551.16 | 41.90 | 47.01 |
| Parameters | 36.90M | 270.85K | 560.77K | 313.07K | 429.83K | 284.14K | 7.30M | 432.75K | 1.35M |
| Training Time(s) | 2.77 | 0.29 | 2.42 | 0.24 | 0.23 | 0.32 | 4.01 | 0.28 | 0.45 |


Fig. 11. Impact of input patch sizes.

*2) Effect of Different Input Patch Sizes:* The input data for GFSFN includes HSI and LiDAR patches. Different sizes of patches imply different spatial scale information of land cover. Therefore, patch size is a key parameter that affects classification performance. Fig. 11 shows the OA of three datasets at different patch sizes $\{9 \times 9, 11 \times 11, 13 \times 13, 15 \times 15\}$.

It can be clearly seen from Fig. 11 that the trend of OA changes in the three datasets is first increasing and then decreasing, reaching the optimal level at patch $11 \times 11$. However, as the patch continues to increase, the classification performance actually decreases. This is because the excessively large patch size brings too much redundant information, which may make it difficult for the model to capture key local features, thereby reducing classification performance. In summary, selecting the appropriate patch size is crucial for improving the classification performance of GFSFN. A patch that is too small may not fully capture the global contextual information in the image, while a patch that is too large may increase the computational burden and cause overfitting problems.

### F. Complexity Analysis

A detailed comparison of the complexity of various methods has been conducted on the Houston 2013 dataset. The evaluation indicators include computational cost, parameter count, and training time per epoch, where computational cost is quantified by floating-point operations per second (FLOPs). It can be clearly seen from Table V that FusAtNet based on attention mechanism and multilayer convolution structure has the highest computational cost and parameter count. The training time of GAMF and GLT is longer compared to other methods, due to the high complexity of GAMF's graph structure and the need for GLT to handle inputs at multiple scales. However, methods such as MFT and HCT demonstrate relatively fast training speed due to their lower parameter count and computational requirements. Overall, the proposed method achieves optimal classification performance while maintaining moderate FLOPs, model parameters, and training time.

## IV. DISCUSSION

### A. Analysis of t-SNE Visualization

To further evaluate the classification performance of the proposed GFSFN, t-SNE visualization comparisons were conducted on the Houston 2013, Trento, and MUUFL datasets using GLT, MFT, HCT, CrossHL, which have higher classification accuracy among the comparison methods, and the GFSFN method proposed in this article. From Figs. 12–14, it can be seen that the clustering effect of the method proposed in this paper is the best. Compared with other methods, the method proposed in this article has smaller intra class distances and better clustering performance on some categories, especially the 8th and 11th categories, on the Houston 2013 dataset. In the Trento dataset, the method proposed in this article has relatively few cases of category confusion, especially in the third category where there is almost no confusion. On the MUUFL dataset, compared to other methods, the proposed method has fewer cases of category confusion in the 5th and 8th classes.

### B. Analysis of Heatmap

In order to more intuitively demonstrate the feature extraction capabilities of the CTEM and S&FA modules, this article presents the visualization results of the proposed module's thermal map, as shown in Figs. 15–17. Corresponding to Cases 3–6 from (a) to (d), cool colors represent low response and warm colors represent high response. Obviously, the addition of CTEM makes the model more sensitive to local complex texture features such as trees, which is particularly evident in the visualization of the Trento and MUUFL datasets. Compared with Case3, increasing S&FA significantly reduces the globally unresponsive regions in the feature map. The combination of CTEM and S&FA achieved the best results, reflecting the effectiveness of the module and the advantages of complementary global and local feature.

### C. Analysis of Frequency Domain Attention

To verify and analyze the effectiveness of the proposed attention module, the classification performance of the proposed module is compared with those of four advanced frequency

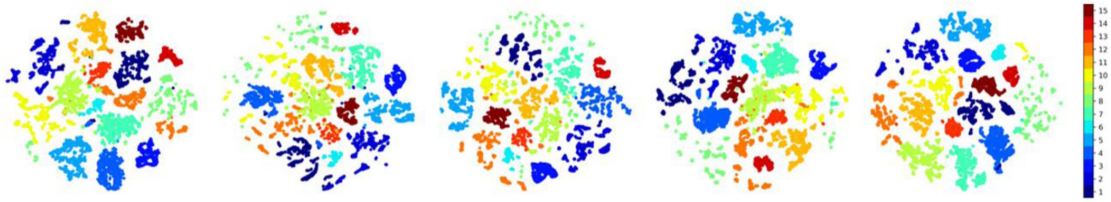



Fig. 12. Comparison of t-SNE visualization results on the Houston 2013 dataset. (a) GLT,(b) MFT,(c) HCT,(d) CrossHL,(e)GFSFN.

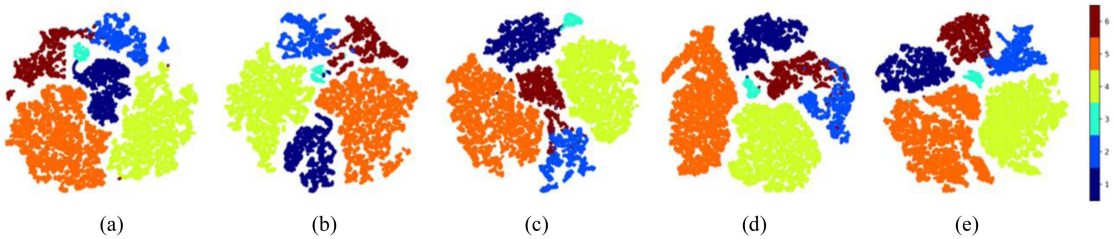



Fig. 13. Comparison of t-SNE visualization results on the Trento dataset. (a) GLT,(b) MFT,(c) HCT,(d) CrossHL,(e) GFSFN.

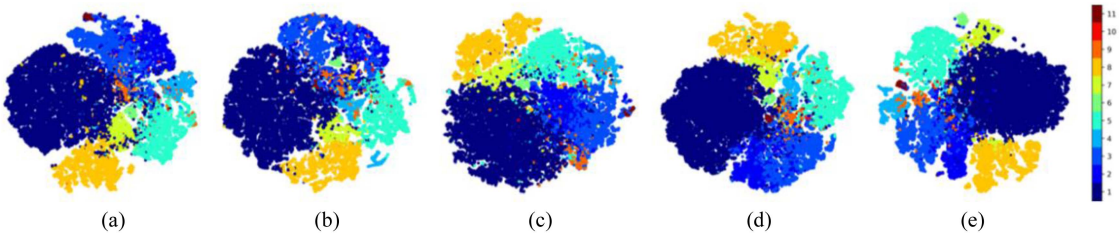



Fig. 14. Comparison of t-SNE visualization results on the MUUFL dataset. (a) GLT,(b) MFT,(c) HCT,(d) CrossHL,(e) GFSFN.

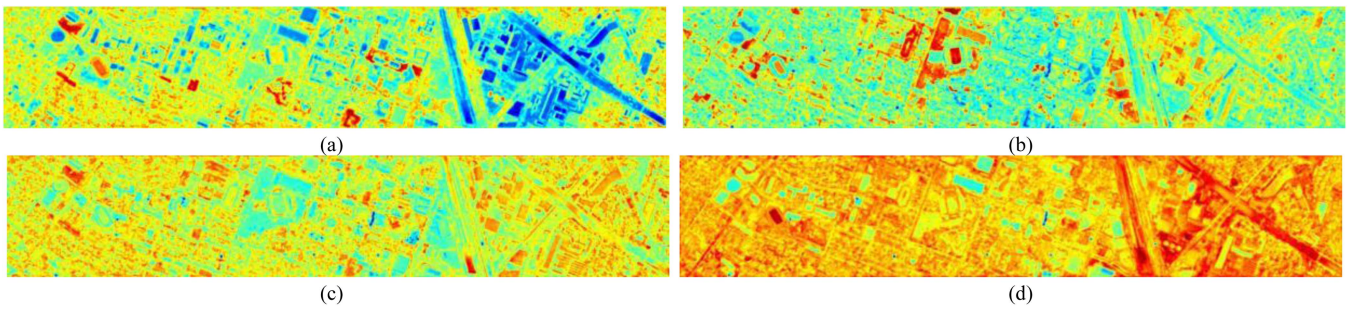



Fig. 15. Comparison of heatmap results of different module on the Houston 2013 dataset. (a) Only CMGF. (b)+CTEM. (c)+S&FA. (d)+CTEM+S&FA.

domain attention modules in the field of image processing. The experimental results are shown in Table VI. Among them, WSA [47] and FSA [48] are based on discrete wavelet transform (DWT), while MCFA [49] and FAM [50] are based on FFT, just like our method. Specifically, WSA is an improvement based on SE attention, which replaces the original pooling operation with DWT, aiming to obtain more critical feature maps by aggregating high and low frequency features. FSA decomposes input features through DWT, performs traditional attention calculations on low-frequency components, and finally restores feature shapes through reverse DWT. In MSFA, the features are first mapped to

TABLE VI
COMPARISON OF OA WITH DIFFERENT FREQUENCY DOMAIN ATTENTION MODULES

| Module | Houston2013 | Trenton | MUUFL |
|---|---|---|---|
| WSA [47] | 96.56 | 99.25 | 85.07 |
| FSA [48] | 95.63 | 98.83 | 84.24 |
| MCFA[49] | 96.45 | 99.19 | 85.29 |
| FAM [50] | 96.01 | 98.91 | 84.94 |
| Ours | **97.23** | **99.59** | **85.96** |

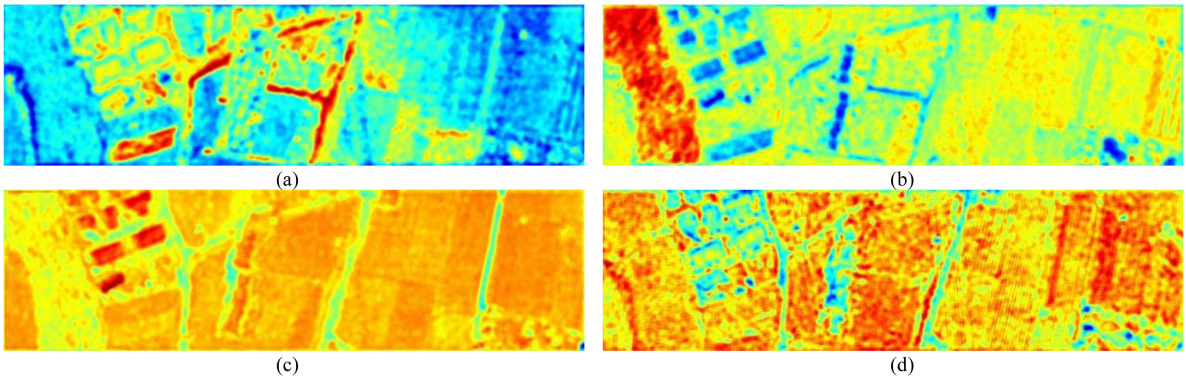



Fig. 16. Comparison of heatmap results of different module on the Trento dataset. (a) Only CMGF. (b)+CTEM. (c)+S&FA. (d)+CTEM+S&FA.

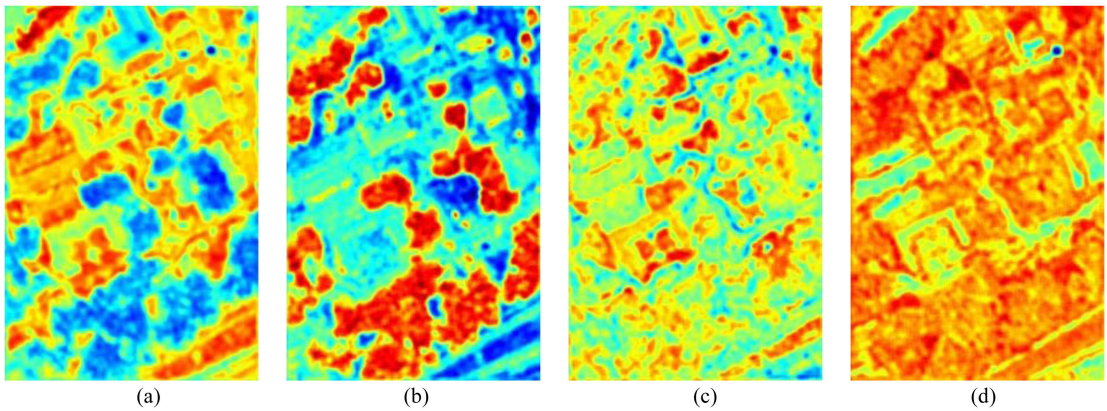



Fig. 17. Comparison of heatmap results of different module on the MUUFL dataset. (a) Only CMGF. (b)+CTEM. (c)+S&FA. (d)+CTEM+S&FA.

the frequency domain through FFT, and the real and imaginary parts are filtered before being transformed back to the spatial domain through IFFT for multiscale convolution operations. FAM performs FFT on the features, performs global filtering and high pass filtering to obtain high-frequency information, and then uses IFFT to map back to the spatial domain and combine the original features obtained through pointwise convolution.

Obviously, overemphasizing the learning of low-frequency or high-frequency components is not conducive to effective feature extraction. Therefore, the OAs of FSA and FAM are relatively low. Other FFT based modules adopt preset filtering operations to process frequency features, which may result in the loss of specific information, while our S&FA constructs attention weights in the frequency domain, and thus can retain and utilize key information more effectively. The excellent classification performance demonstrated by WSA also validates the effectiveness of combining DWT and attention. Overall, the attention module that combines spatial and frequency domain feature processing can bring better feature learning ability and classification performance, such as MCFA and our proposed S&FA.

### *D. Analysis of Sampling Strategy*

There are generally two strategies to divide the training set and the testing set: random sampling and disjoint sampling. For random sampling, data from each category is randomly assigned to the training or testing set according to a predetermined ratio or specific quantity. For disjoint sampling, the dataset is strictly divided into non overlapping training and testing sets. The experimental results in this paper are based on random sampling. To compare and analyze the performance differences of the models under these two different sampling techniques, some standard disjoint sampling experiments have been conducted on the Houston 2013 dataset. The experimental results are shown in Fig. 18. In Fig. 18, the line represents the experimental results with random sampling as shown in Table I, and the bar graph represents the experimental results with disjoint sampling. In addition, three different colors represent OA, AA, and Kappa, respectively.

It can be clearly observed from Fig. 18 that all methods have lower results than random sampling when conducting experiments using disjoint strategies. Specifically, the performance difference between FusAtNet and GAMF is particularly prominent, while other methods also show a certain degree of performance degradation after implementing disjoint sampling. FusAtNet relies on its complex convolutional structure, while GAMF is based on graph attention techniques, both of which exhibit strong spatial dependence. Therefore, their performance under disjoint sampling is not satisfactory. Although random sampling helps improve the robustness of the model, compared to disjoint sampling, the random sampling strategy may also

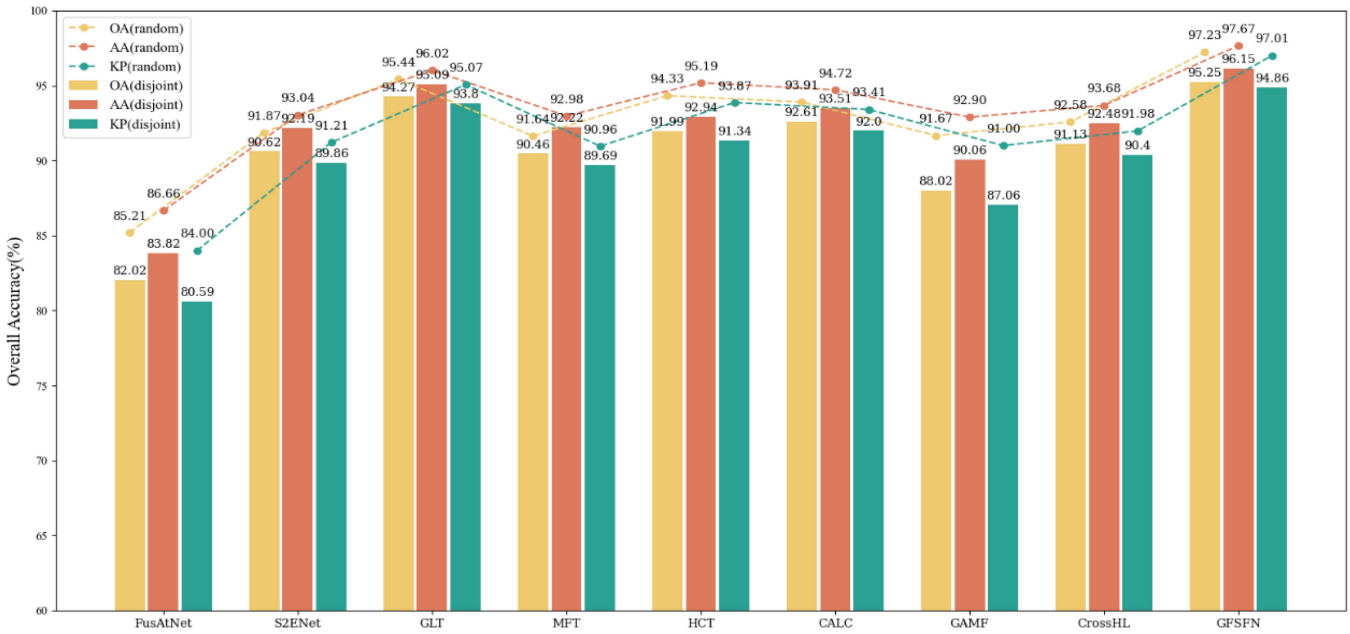



Fig. 18. Experimental results of different sampling strategies on the Houston 2013 dataset.

come with the risk of information leakage. It is worth emphasizing that regardless of which sampling method is adopted, random sampling or disjoint sampling, the method proposed in this paper has always achieved the best classification performance, which fully demonstrates the effectiveness of the proposed method.

## V. Conclusion

In this article, we propose a GFSFN to address the joint classification problem of hyperspectral and radar data. The proposed GFSFN consists of three core modules: CMGF, S&FA, and CTEM. To reduce the differences in heterogeneous data and effectively fuse multimodal features, CMGF is proposed to fuse the dilated convolution features of two modalities into spatial features at the same channel depth, and then adaptively fuse the joint features at different depths through a gating mechanism. To address the neglect of frequency domain features in current joint classification, a S&FA is proposed, which constructs an attention weight matrix in the frequency domain to learn and model global features, especially fine features, through the interaction between frequency domain and spatial domain features. To enhance the extraction of local fine features, a convolution based CTEM has been proposed, further improving the accuracy of classification. Experiments on three common datasets show that the GFSFN proposed in this article has excellent feature learning ability compared to other advanced classification methods.

In future work, we will explore deeper feature representations in the frequency domain and effectively combine them with spatial domain features. At the same time, we will further consider how to improve the model to enhance the local global correlation of features, to further improve the classification performance and training efficiency of the model.

## References

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[2] A. Ibrahim et al., "Atmospheric correction for hyperspectral ocean color retrieval with application to the hyperspectral imager for the coastal ocean (HICO)," *Remote Sens. Environ.*, vol. 204, pp. 60–75, Jan. 2018.

[3] A. Backhaus, F. Bollenbeck, and U. Seiffert, "Robust classification of the nutrition state in crop plants by hyperspectral imaging and artificial neural networks," in *Proc. 3rd Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2011, pp. 1–4.

[4] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Feb. 2014.

[5] R. Idris, Z. A. Latif, J. Jaafar, N. M. Rani, and F. Yunus, "Quantitative assessment of LiDAR dataset for topographic maps revision," in *Proc. Int. Conf. System Eng. Technol.*, 2012, pp. 1–4.

[6] S. Li, Q. Liu, Z. Li, Z. Qi, L. Si, and N. Wang, "Forest canopy gap dynamics based on Time-Series of airborne lidar data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 6119–6121.

[7] W. Li, W. Zhou, Y. M. Wang, C. Shen, X. Zhang, and X. Li, "Meteorological radar fault diagnosis based on deep learning," in *Proc. Int. Conf. Meteorol. Observ.*, 2019, pp. 1–4.

[8] Z. Huang, H. Qi, C. Kang, Y. Su, and Y. Liu, "An ensemble learning approach for urban land use mapping based on remote sensing imagery and social sensing data," *Remote Sens*, vol. 12, no. 19, Oct. 2020, Art. no. 3254.

[9] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and LiDAR remote sensing data for classification of complex forest areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1416–1427, May 2008.

[10] P. Ghamisi, R. Souza, J. A. Benediktsson, X. Xiao, and R. A. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, Oct. 2016.

[11] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and LiDAR data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.

[12] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and LiDAR data," *Int. J. Image Data Fusion*, vol. 6, no. 3, pp. 189–215, Jul. 2015.
[13] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.
[14] 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.
[15] 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.
[16] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5500205.
[17] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514116.
[18] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506812.
[19] W. Yu, H. Huang, M. Zhang, Y. Shen, and G. Shen, "Shadow maskdriven multimodal intrinsic image decomposition for hyperspectral and LiDAR data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5525915.
[20] C. Shi, X. Zhang, L. Wang, and Z. Jin, "A lightweight convolution neural network based on joint features for Remote sensing scene image classification," *Int. J. Remote Sens.*, vol. 44, no. 21, pp. 6615–6641, 2023.
[21] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
[22] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[23] 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.
[24] J. Li, Y. Liu, R. Song, Y. Li, K. Han, and Q. Du, "Sal2rn: A spatial–spectral salient reinforcement network for hyperspectral and lidar data fusion classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500114.
[25] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on expansion convolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528316.
[26] C. Shi, J. Chen, and L. Wang, "Hyperspectral image classification based on a novel Lush multi-layer feature fusion bias network," *Expert Syst. Appl.*, vol. 247, Aug. 2024, Art. no. 123155.
[27] W. Yu, L. Gao, H. Huang, Y. Shen, and G. Shen, "HI2D2FNet: Hyperspectral intrinsic image decomposition guided data fusion network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5521715.
[28] W. Yu, L. Gao, H. Huang, Y. Shen, and G. Shen, "PID-HLfusion: Pluggable progressive illumination driven hyperspectral and LiDAR data fusion considering crossmodal geometric structures," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2529316.
[29] H. Gong, "Implements of transformer in NLP and DKT," in *Proc. 4th Int. Conf. Artif. Intell. Adv. Manuf.*, 2022, pp. 805–807.
[30] F. Yan, B. Yan, and M. Pei, "Dual transformer encoder model for medical image classification," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 690–694.
[31] C. Shi, S. Yue, and L. Wang, "Attention head interactive dual attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5523720.
[32] C. Shi, H. Wu, and L. Wang, "A feature complementary attention network based on adaptive knowledge filtering for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527219.
[33] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for hsi and lidar data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.
[34] W. Yu, H. Huang, and G. Shen, "Deep spectral–spatial feature fusionbased multiscale adaptable attention network for hyperspectral feature extraction," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 5500813.
[35] T. Song, Z. Zeng, C. Gao, H. Chen, and J. Li, "Joint classification of hyperspectral and lidar data using height information guided hierarchical fusion-and-separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505315.
[36] 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, 2022, Art. no. 6504205.
[37] H. Gao et al., "Interactive enhanced network based on multihead self-attention and graph convolution for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5533716.
[38] J. Yang et al., "LiDAR-guided cross-attention fusion for hyperspectral band selection and image classification," *IEEE Trans. Geosci. Remote Sens*, vol. 62, 2024, Art. no. 5515815.
[39] H. Shi, G. Cao, Y. Zhang, Z. Ge, Y. Liu, and D. Yang, "F3Net:Fast Fourier filter network for hyperspectral image classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–18, 2023.
[40] H. Shi, Y. Zhang, G. Cao, and D. Yang, "MHCFormer: Multiscale hierarchical Conv-Aided fourierformer for hyperspectral image classification," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 5501115.
[41] S. Mohla, S. Pande, B. Banerjee, and Subhasis Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 92–93.
[42] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.
[43] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and lidar data using a hierarchicalcnn and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500716.
[44] 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.
[45] J. Cai et al., "A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123587.
[46] 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.
[47] Y. Yang et al., "Dual wavelet attention networks for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1899–1910, Apr. 2023.
[48] F. Zhang, A. Panahi, and G. Gao, "FsaNet: Frequency self-attention for semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 4757–4772, 2023.
[49] C. Yu et al., "Frequency-temporal attention network for remote sensing imagery change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5005305.
[50] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G, Carneiro, and T.-T. Do, "Frequency attention for knowledge distillation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 2277–2286.

**Cuiping Shi** (Member, IEEE) received the M.S. degree in signal and information processing from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

From 2017 to 2020, she was a Postdoctoral Researcher with the College of Information and Communications Engineering, Harbin Engineering University, Harbin. She is currently a Professor with the Department of Communication Engineering, Qiqihar University, Qiqihar, China. Since 2024, she was with the College of Information Engineering, Huzhou University, Huzhou, China. She has authored or coauthored two academic books about remote sensing image processing and more than 90 papers in journals and conference proceedings. Her research interests include remote sensing image processing pattern recognition and machine learning.

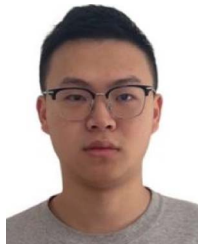
Dr. Shi was the recipient of the nomination award of Excellent Doctoral Dissertation of HIT in 2016 for her doctoral dissertation.

**Zhipeng Zhong** received the bachelor's degree in computer science and technology from Taizhou University, Linhai, China, in 2022. He is currently working toward the master's degree in computer science and technology with Huzhou University, Huzhou, China.

His research interests include hyperspectral image processing and machine learning.

**Shihang Ding** received the bachelor's degree in bioengineering from the Henan University of Technology, Zhengzhou, China, in 2022. He is currently working toward the master's degree in computer science and technology with Huzhou University, Huzhou, China.

His research interests include hyperspectral image processing and machine learning.



**Yeqi Lei** received the bachelor's degree in food science and engineering from Jiangsu University, Zhenjiang, China, in 2023. He is currently working toward the master's degree in electronic information with Huzhou University, Huzhou, China.

His research interests include hyperspectral image processing and machine learning.

**Liguo Wang** (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a postdoctoral research position from Harbin Engineering University. Since 2020, he has been working with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China. His main research interests include remote sensing image processing. He has published two books about hyperspectral image processing and more than 130 papers in journals and conference proceedings.

**Zhan Jin**, photograph and biography not available at the time of publication.