Multiscale Spectral–Spatial Attention Residual Fusion Network for Multisource Remote Sensing Data Classification

Xu Wang, Gang Liu, Ke Li, *Graduate Student Member, IEEE*, Min Dang, *Graduate Student Member, IEEE*, Di Wang, *Member, IEEE*, Zili Wu, and Rong Pan

***Abstract*—The joint of multisource remote sensing (RS) data for land cover classification has become a popular research topic. Although studies have shown that the fusion of multisource data can improve the accuracy of classification, the current limitation lies in the inadequate exploitation of information, resulting in spectral confusion categories overlapping, and varying visual differences within the same category. To address these problems, this article proposes a multiscale spectral–spatial attention residual fusion network (MSSARFNet) that aims to enhance the classification performance of multisource RS data through effective spectral–spatial feature extraction and fusion. Specifically, three modules are designed: the multiscale spectral attention residual module (MSpeARM), the multiscale spatial attention residual module (MSpaARM), and multiscale convolution fusion (MCF) module. First, we divide the channels of the features into multiple paths and apply spectral–spatial attention mechanisms on each path. To further enhance connectivity, convolutional operation is utilized to establish connections between different paths, thus forming MSpeARM and MSpaARM. The MSpeARM suppresses redundant features and enhances effective features along the channel dimension to better differentiate spectral-confused categories. The MSpaARM highlights that different visual patterns of objects in the same category can mutually reinforce each other by weighting all positional features, regardless of their spatial differences. Second, to fuse these two sets of features, the MCF module is designed to learn multilevel semantic features and enhance fusion at a granular level. Experimental evaluations on three RS datasets demonstrate that the proposed method achieves excellent classification performance, indicating the effectiveness of MSSARFNet.**

Manuscript received 24 January 2024; revised 29 February 2024; accepted 13 March 2024. Date of publication 20 March 2024; date of current version 4 April 2024. This work was supported in part by the Key Research and Development Program of Shaanxi Province, China under Grant 2023-YBGY-205, in part by the Natural Science Basic Research Program of Shaanxi, China under Grant 2024JC-ZDXM-40, in part by the Key Research and Development Program of Shaanxi, China under Grant 2024GXYBXM-039, and in part by the Innovation Capability Support Program of Shaanxi, China, under Grant 2023-CX-TD-08. *(Corresponding author: Gang Liu.)*

Xu Wang, Ke Li, Min Dang, Di Wang, Zili Wu, and Rong Pan are with the Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: wangxu0210@stu.xidian.edu.cn; like0413@stu.xidian.edu.cn; dangmin@stu.xidian.edu.cn; wangdi@xidian.edu.cn; zlwu@xidian.edu.cn; rpan@xidian.edu.cn).

Gang Liu is with the School of Computer Science and Technology, Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, Xidian University, Xi'an 710071, China, and also with the Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China (e-mail: gliu@xidian.edu.cn).

The source code of the proposed method will be available at https://github.com/wangxu0210/MSSARFNet.

Digital Object Identifier 10.1109/JSTARS.2024.3379579

***Index Terms*—Multiscale spectral–spatial attention residual fusion network (MSSARFNet), multisource data.**

I. Introduction

With the continuous development of remote sensing (RS) technology, different types of RS data can be acquired in the same observation scene. In the past few years, these data have presented new methods and challenges in various RS fields, including land use and land cover classification [1], [2], hyperspectral anomaly detection [3], semantic segmentation [4], and super-resolution [5]. Recently, Hong et al. [6] proposed SpectralGPT, a model specifically designed for handling RS data, which has shown great potential in tasks such as scene classification and semantic segmentation, offering new perspectives and avenues for advancement in the RS community. In this article, our main focus is on land cover classification, a crucial and challenging task in RS. This task holds increasing importance in various domains such as urban planning and precision agriculture.

In reality, hyperspectral image (HSI) has been widely studied in land cover classification tasks due to its rich spatial and spectral information and its ability to provide comprehensive spectral information about the ground. For years, researchers have been working on developing more efficient feature extractors for HSI: for instance, Roy et al. [7] combined the morphological operation and deep learning (DL) and proposed a morphological CNN, which presented powerful nonlinear transformations for feature extraction. Ding et al. [8] used an ARMA filter to extract robust HSI features for classification. Furthermore, Ding et al. [9] proposed a new method of combining graph convolution with adaptive filters that learns spatial and spectral features to improve the classification performance. These well-designed feature extractors enable the potential of HSI in classification tasks. However, different types of ground objects usually have similar spectral curves. Furthermore, differences in regional distribution can lead to different spectral curves even within the same type of ground object. Therefore, relying solely on HSI data for complex scene classification tasks is difficult. Unlike HSI data, the light detection and ranging (LiDAR) can provide elevation distribution information [10]. However, the existing limitation lies in the insufficient utilization of spectral information during the extraction and fusion process of multisource data, resulting in spectral confusion, category overlap, and different

visual differences for the same category due to insufficient spatial information. This poses a challenge for multisource land classification.

Recently, to fully leverage the HSI and LiDAR data, numerous studies have proposed methods primarily focused on extracting multiple features from them. The extracted features are subsequently fused to achieve accurate land cover classification.

To effectively utilize the diverse information from multisource data and perform dimension reduction [12], massive feature extraction methods have been proposed. Traditional machine learning algorithms [12], [13], [14] and convolutional neural networks (CNNs) are the most widely used feature extraction methods. Support vector machines (SVM) [13] and Random Forest (RF) [14] are frequently employed in the early works. As for CNNs, the ability to automatically extract deep abstract features has led to their widespread use as a feature extraction method. Chen et al. [15] proposed a dual-stream model in which depth data were extracted from the LiDAR using a 2-D CNN, while spectral–spatial features from HSI were extracted using a 3-D CNN in another stream. Xu et al. [16] designed a complex CNN architecture called two-branch CNN (TBCNN) for multisource classification. The TBCNN consists of two branches: an HSI branch for spatial and spectral information extraction from HSI data, and a LiDAR branch with especially designed cascade blocks for LiDAR feature extraction. Hong et al. [17] proposed a shared and specific feature learning model that can extract more useful information from multisource RS data. Although there may exists some redundancy in the extracted features from multisource data, the feature fusion procedure lacks refinement, resulting in an insufficient utilization of the information correlation between data sources. For the multisource RS data classification, it is critical to effectively utilize the correlation and complementary information among different data sources.

To effectively fuse the features extracted from multisource data, several feature fusion methods have been proposed. Liao et al. [18] improved the morphological contour extraction for feature fusion and classification of HSI and LiDAR data by applying a graph-based fusion strategy. Rasti et al. [19] used total variation component analysis to fuse feature. To further enhance the classification accuracy, Feng et al. [20] proposed an adaptive HSI-LiDAR fusion method, which effectively combines HSI and LiDAR features in a more reasonable and natural manner. Hang et al. [21] introduced a coupled CNN and developed simultaneous feature-level fusion and decision-level fusion strategies. These strategies were designed to enhance the classification performance by effectively fusing features and making decisions in a coordinated way. Wu et al. [22] created a cross-channel reconstruction module to enhance the feature fusion representations of various RS data. Zhang et al. [23] utilized Gram matrices to enhance the preservation of complementary information from multisource data in a TBCNN fused with HSI and LiDAR data. While the aforementioned methods have significantly improved the classification performance, they still suffer from spectral information redundancy during the fusion process. This makes it difficult to distinguish certain confusing objects, and the insufficient utilization of spatial information leads to different visual differences within the same category.

Based on the aforementioned discussion, our motivation is to allow models to mine and focus on key information in multisource data from both spectral and spatial dimensions. Thus, a multiscale spectral–spatial attention residual fusion network (MSSARFNet) is proposed. First, shallow features are extracted from multisource data using a two-branch HSI and LiDAR feature representation network. Then, two attention modules are designed, one being the multiscale spectral attention residual module (MSpeARM) and the other being the multiscale spatial attention residual module (MSpaARM). The MSpeARM simulates the interdependence among channels at multiple scales, emphasizing the independence of channels. In this way, it suppresses redundant features and enhances effective features along the channel dimension to better differentiate spectral confusion categories. The MSpaARM highlights that different visual patterns of objects in the same category can mutually reinforce each other by weighting all positional features, regardless of their spatial differences in shape and distribution. Next, to integrate the features of the two types, a multiscale convolutional fusion (MCF) module is designed. This module aggregates contextual information from spectral–spatial features and fully exploits multiscale spectral–spatial information for effective fusion. Finally, the fused features are fed into the classification module to obtain the final classification results. Experimental evaluations on three multisource RS datasets demonstrate that the proposed method achieves excellent classification performance, indicating the effectiveness of the proposed MSSARFNet.

The main contributions of our work can be summarized as follows.

1) Two multiscale attention modules are proposed, which split the feature channels into multiple paths. These paths incorporate spectral and spatial attention mechanisms in both spectral and spatial dimensions. By applying these mechanisms separately to each path, the model is guided to emphasize the crucial information in the image through attention weights. By analyzing the feature information of each path and correlating each path, the multiscale spectral–spatial information in multisource data can be fully exploited, which is more suitable for RS scene classification.

2) To further integrate the extracted multiscale spectral–spatial information, this article introduces the MCF module. The MCF module adopts a multibranch approach to extract and fuse features at different scales, enhancing the spectral–spatial representation capacity and enabling the model to comprehensively understand multiscale spectral–spatial information. Through effective fusion, the overall classification performance is improved.

The rest of this article is organized as follows. In Section II, related work is introduced. Section III describes the proposed method. Extensive experiments and analyses are presented in Section IV. Finally, Section V concludes this article.

II. Related Work

A. CNNs-Based Multisource Classification

With the successful introduction and rapid development of DL methods, it has achieved outstanding performance in multisource representation [24], [25]. DL methods can effectively

learn spectral and spatial features from multisource data, demonstrating significant potential in the joint classification of multisource data. Hong et al. [26] developed a deep encoder and decoder network for pixel-level classification of multimodal RS data. Hong et al. [27] also proposed a unified multimodal DL framework that assembles pixel-level labeling guided by a fully connected (FC) design and spectral–spatial joint classification with CNNs-dominated architecture. To further enhance the joint classification accuracy, Zhao et al. [28] utilized Octave convolution to reduce redundant features in the multisource data. This was followed by fractional Gabor convolution to capture multiscale, multidirectional spectral–spatial features and effectively integrate them. Wang et al. [29] developed a method to combine spectral information from HSI with LiDAR data using spectral–spatial interconductivity modules and adaptive multiscale mutual learning. Furthermore, Gao et al. [30] introduced an adaptive multiscale spectral–spatial enhancement network classification method that combines pairwise ensemble operators, multibranch extraction, and adaptive feature fusion to incorporate spectral and spatial information from HSIs and LiDAR data. Dong et al. [31] utilized a method of dictionary-based distribution alignment to facilitate the complementary integration of multisource data, leading to improved accuracy in classification.

B. Attention-Based Classification: From Single Source to Multisource

Recently, the attention mechanism has shown significant performance gains for different visual tasks [32], [33], [34]. It allows a model to better focus on important regions and suppress irrelevant features through self-supervised learning paradigms. Attention have also been employed to RS classification by some studies [35], [36]. Li et al. [37] developed a mechanism network comprised of channel and spatial attention blocks. This enables network to enhance and optimize the retrieved feature maps. Haut et al. [38] created a novel attention-driven classification network that enhances the feature extraction capability by introducing residual learning [39] and the dual data path attention modules as fundamental building blocks. Zhu et al. [40] investigated the attention mechanisms in spectral and spatial information that selectively captures useful features in both spectral and spatial dimensions to enhance the classification performance. Mohla at el. [41] used the self-attention mechanism to highlight spectral features of HSI and the cross-attention mechanism to accentuate the spatial features of HSI via a LiDAR-derived attention map. Li at el. [42] integrates spectral and spatial attention components to facilitate the interaction between HSI and LiDAR data. The method leveraged the information from one modality to enhance the performance of the other, improving the overall classification accuracy.

Furthermore, some studies have proposed multihead attention mechanisms that incorporate spectral–spatial information. Hong et al. [43] designed a spectral transformer model that aims to capture local sequential features in the spectral domain. Sun et al. [44] proposed a novel spectral–spatial feature tokenization transformer (SSFTT) for capturing high-level semantic information and spectral–spatial characteristics. Ding et al. [45] designed a two-layer 1-D CNN spectral transformer mechanism to extract the spectral features of images, with which the spectral features can be acquired automatically. Roy et al. [46] proposed a new multimodal fusion transformer (MFT) network for HSI and LiDAR joint classification, which includes a multihead cross-patch attention. This method, however, fell short in fully integrating the relevant information from both data modalities. To address this limitation, Zhao et al. [47] designed a novel network that joints convolution and transformer to extract spatial–spectral information and achieves effective fusion. Li et al. [48] proposed a unified framework that incorporated a multihead cross-modal attention mechanism to capture the interplay between multisource data and aggregate contextual information. Yao et al. [49] proposed a general multimodal transformer framework that designed a hybrid spatial vision transformer backbone implemented with both self-attention and cross-modality attention mechanisms for better information fusion in classifying multimodal RS data. Furthermore, Zhao et al. [50] designed a fractional Fourier image transformer (FrFT), which was used as a backbone network to extract multimodal global and local contextual information extraction.

Throughout the aforementioned related works, although significant improvements have been made in the evaluation of multisource classification performance, some methods still do not fully utilize spectral or spatial information, leading to information redundancy, spectral confusion, and different visual modalities within the same category. In addition, most methods treat the feature channel as a whole and do not consider dividing it into multiple paths, allowing each path to learn, and correlating the information obtained from each path using operations such as convolution. Therefore, we divide the feature channel into multiple paths and apply spectral and spatial attention mechanisms to each path to fully learn the spectral–spatial information.

III. Methodology

The framework of the proposed MSSARFNet is illustrated in Fig. 1. It consists of six parts: HSI and LiDAR data preprocessing, HSI and LiDAR feature learning via CNN, MSpeARM, MSpaARM, MCF module, and classification module.

A. HSI and LiDAR Data Preprocessing

Given an HSI as $X_H \in \mathbb{R}^{M\times N\times B}$ and a LiDAR as $X_L \in \mathbb{R}^{M\times N}$, where M and N are the width and the height of the spatial region respectively, and B is the number of HSI spectral bands. There are typically massive spectral bands that provide useful information. However, this leads to an increase in computation cycles. To reduce the spectral dimension, PCA is used to extract the first b principal components of X_H, maintaining the spatial dimensions but lowering the number of bands from B to b. After PCA, the HSI data X_H are transformed into $X_H^{\text{pca}} \in \mathbb{R}^{M\times N\times b}$.

Next, for each pixel, 3-D and 2-D patch extraction are performed to obtain a small patch cube $X_H^P \in \mathbb{R}^{p\times p\times b}$ and a small

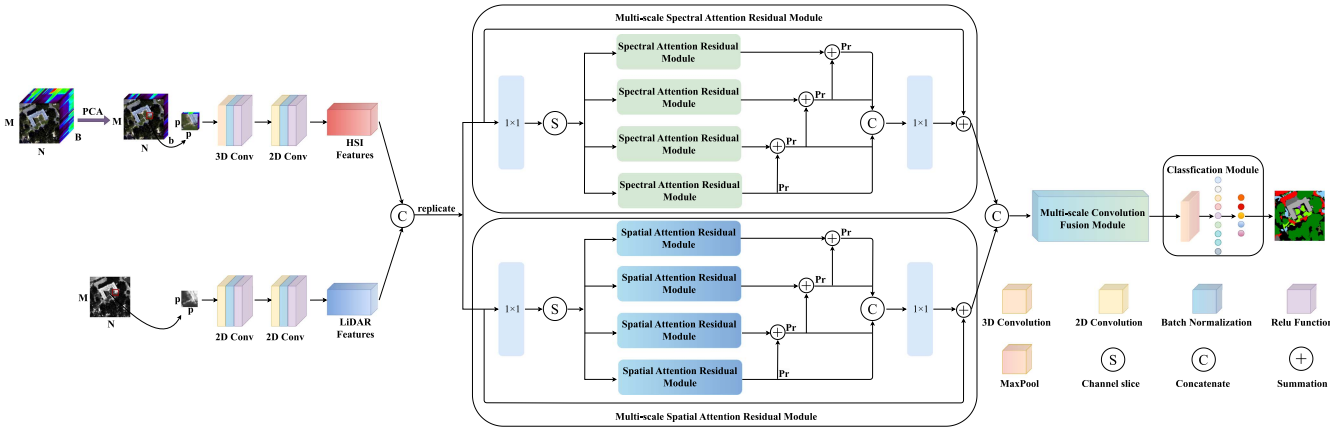


Fig. 1. Overview of the proposed MSSARFNet for multisource classification. Here, "replicate" represents a copy of the shallow information data obtained from the two-branch convolution network. Pr is the combination of the convolution operation. The details of SpeARM, SpaARM, and MCF module can be seen in Figs. 2–4.

patch $X_L^P \in \mathbb{R}^{p\times p}$, where $p \times p$ is the patch size. For edge pixels, the padding operation is applied to these pixels with a size of $(p-1)/2$. Based on the ground-truth map, the samples are divided into a training set and a test set after removing the pixel blocks with a label of zero.

B. HSI and LiDAR Feature Learning Via CNN

The CNN can automatically extract contextual and high-level abstract features, and demonstrate powerful modeling capabilities, which results in excellent performance in the HSI classification. Furthermore, the HybridSN proposed by Roy et al. [51] demonstrates that 3-D CNN promotes combined spatial–spectral feature representation from a stack of spectral bands. After the 3-D CNN, a 2-D CNN is used to further learn more abstract spatial representations. Therefore, spectral–spatial information and elevation information are extracted from HSI data and LiDAR data, respectively, using a TBCNN in the proposed framework.

As shown in Fig. 1, for the HSI data, sequential the Conv3D and Conv2D layers are used. The HSI cube X_H^P of size $p \times p \times b$ is first unsqueezed into shape $1 \times p \times p \times b$, subsequently fed into the Conv3D layer for training. In the Conv3D layer, a convolution with eight kernels of spatial size $3 \times 3 \times 3$ is applied. Then, the 3-D data are converted to 2-D data for subsequent 2-D convolution. The Conv2D layer uses a convolution with 64 kernels of spatial size 3×3. After the operation, feature maps can be generated

$$X_H^{\text{out}} = \text{Conv2D}(\text{Reshape}(\text{Conv3D}(X_H^P))) \tag{1}$$

where the shapes of the output feature maps after the Conv3D layer and the Conv2D layer are $8 \times (p-2) \times (p-2) \times (b-8)$ and $(p-4) \times (p-4) \times 64$, respectively.

Unlike the HSI processing, the two Conv2D layers are used to extract LiDAR elevation features. In the two layers, the convolution with 16 and 64 kernels of spatial size 3×3 are applied, respectively,

$$X_L^{\text{out}} = \text{Conv2D}(\text{Conv2D}(X_L^P)) \tag{2}$$

where the shapes of the output feature maps after the sequential Conv2D layer are $(p-2) \times (p-2) \times 64$ and $(p-4) \times (p-4) \times 64$, respectively. To regularize and speed up the training process, a batch normalization layer and rectified linear units (ReLU) function are consecutively applied after each convolutional layer.

C. Multiscale Spectral Attention Residual Module (MSpeARM)

The redundancy of spectral information makes it difficult to differentiate some confusing categories in multisource classification. Therefore, the MSpeARM containing m groups spectral attention residual modules (SpeARMs) (as shown in Fig. 1) was designed to refine these spectral features to selectively emphasize the different spectral channel. By highlighting or suppressing some channel mappings, we can capture the spectral dependencies between the multiple channels and enhance the feature representation of specific categories.

First, the feature map of the input is defined for the MSpeARM, which is derived from the shallow information obtained by TBCNN as follows:

$$\begin{aligned} F_{\text{in}}^{\text{Spe}} &= \text{Concat}(X_H^{\text{out}}, X_L^{\text{out}}) \\ F^{\text{Spe}} &= \text{Conv}_{1\times1}(F_{\text{in}}^{\text{Spe}}) \end{aligned} \tag{3}$$

where $F^{\text{Spe}} \in \mathbb{R}^{c\times s\times s}$, c represents the number of multisource data channels, s is the size of the feature map after TBCNN, $\text{Concat}(\cdot)$ represents the joint operation, and $\text{Conv}_{1\times1}(\cdot)$ is the 2-D convolutional layer with kernel size 1×1.

Subsequently, the F^{Spe} is first sliced into m groups of feature maps: F^{Spe_1}, F^{Spe_2},..., F^{Spe_m}, where m is a power of 2 and $m \in \{1, 2, 4, \ldots, c\}$. Second, to construct the spectral attention maps, each group of feature maps is sent through the SpeARM. Finally, the multiscale spectral feature maps may be generated using these spectral attention maps $F_{\text{out}}^{\text{Spe}} \in \mathbb{R}^{c\times s\times s}$ through (10). The kth group $F^{\text{Spe}_k} \in \mathbb{R}^{c/m\times s\times s}$ is used to demonstrate the SpeARM, where $k \in [1, m]$. In the SpeARM [as shown in

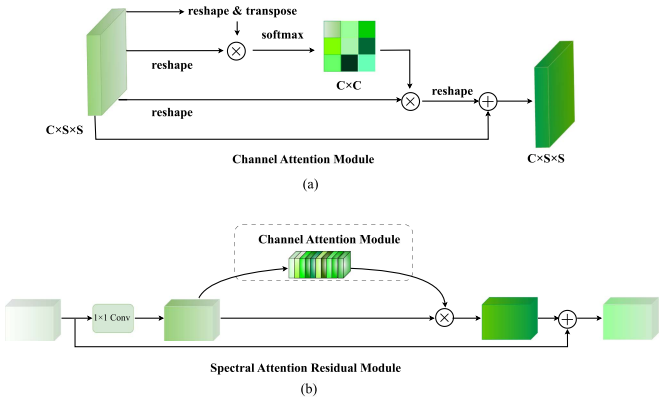


Fig. 2. Illustration of the SpeARM.

Fig. 2(b)], the input feature map $F^{\mathrm{Spe}'_k} \in \mathbb{R}^{c/m\times s\times s}$ is first executed with a 1×1 convolution operation and an ReLU activation operation.

$$F^{\mathrm{Spe}'_k} = \mathrm{ReLU}(\mathrm{Conv}_{1\times 1}(F^{\mathrm{Spe}_k})) \tag{4}$$

where $\mathrm{ReLU}(\cdot)$ represents the activation function.

Inspired by Fu at el. [52], we introduce the channel attention mechanism [as shown in Fig. 2(a)] to model the interdependence across spectra feature, thereby capturing the interdependencies among spectral information in the multisource data. Specifically, put $F^{\mathrm{Spe}'_k} \in \mathbb{R}^{c/m\times s\times s}$ into $F^{\mathrm{Spe}'_k} \in \mathbb{R}^{c/m\times g}$, where $g = s \times s$. Thus, the channel attention map $A^{\mathrm{Spe}_k} \in \mathbb{R}^{s\times s}$ can be formulated as

$$A^{\mathrm{Spe}_k}_{ji} = \frac{\exp\left(F^{\mathrm{Spe}'_k}_i \times F^{\mathrm{Spe}'_k}_j\right)}{\sum_{i=1}^{c}\exp\left(F^{\mathrm{Spe}'_k}_i \times F^{\mathrm{Spe}'_k}_j\right)} \tag{5}$$

where $A^{\mathrm{Spe}_k}_{ji}$ represents the influence of the ith channel on the jth channel. $F^{\mathrm{Spe}'_k}_i$ represents the ith feature map of the $F^{\mathrm{Spe}'_k}$.

Then, the results of matrix multiplication between A^{Spe_k} and $F^{\mathrm{Spe}'_k}$ are reshaped into $\mathbb{R}^{c/m\times s\times s}$. The revised results are then weighted using a learnable parameter λ. To maintain the original feature information of $F^{\mathrm{Spe}'_k}$, a sum operation is performed between the weighted results and $F^{\mathrm{Spe}'_k}$ to get the kth group spectral feature map $F^{\mathrm{Spe}''_k} \in \mathbb{R}^{c/m\times s\times s}$ as follows:

$$F^{\mathrm{Spe}''_k}_j = \lambda\sum_{i=1}^{c}(A^{\mathrm{Spe}_k}_{ji}) + F^{\mathrm{Spe}'_k}_j. \tag{6}$$

To further improve the representation of the spectral information as well as the whole network, the information obtained from the channel attention mechanism is multiplied as weights with the original features $F^{\mathrm{Spe}'_k}$. And the residual links are introduced to prevent gradient explosion and network degradation as follows:

$$F^{\mathrm{Spe}_k}_{\mathrm{out}} = \sigma(F^{\mathrm{Spe}''_k}) \times F^{\mathrm{Spe}'_k} + F^{\mathrm{Spe}'_k} \tag{7}$$

where $\sigma(\cdot)$ represents sigmiod function.

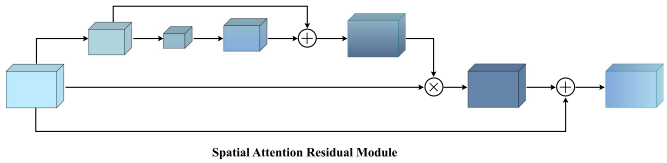


Fig. 3. Illustration of the SpaARM.

To construct spectral features F with a size of $c/m \times s \times s$ from m-group spectral maps, the following operation is adopted:

$$F^i = \begin{cases} \Pr(F^{\mathrm{Spe}_i}_{\mathrm{out}}) & i = 1 \\ \Pr(F^{\mathrm{Spe}_i}_{\mathrm{out}} + F^{\mathrm{Spe}_{i-1}}_{\mathrm{out}}) & i = 2 \\ \Pr(F^{\mathrm{Spe}_i}_{\mathrm{out}} + F_{i-1}) & 2 < i \le m \end{cases} \tag{8}$$

where $\Pr(\cdot)$ represents the convolution operation, which is intended to relate the information obtained between each group.

Then, to form the new feature map $F_{\mathrm{out}} \in \mathbb{R}^{c\times s\times s}$, all the features F^i are connected, and a 1×1 convolution is used to enhance the correlation of feature for subsequent summation with the original feature map $F^{\mathrm{Spe}}_{\mathrm{in}}$:

$$F_{\mathrm{out}} = \mathrm{Conv}_{1\times 1}(\mathrm{Concat}(F^1, F^2, \ldots, F^m)). \tag{9}$$

Finally, the feature map that contains multiscale spectral information $F^{\mathrm{Spe}}_{\mathrm{out}} \in \mathbb{R}^{c\times s\times s}$ is obtained as follows:

$$F^{\mathrm{Spe}}_{\mathrm{out}} = F_{\mathrm{out}} + F^{\mathrm{Spe}}_{\mathrm{in}}. \tag{10}$$

D. Multiscale Spatial Attention Residual Module (MSpaARM)

For the multisource data, different locations may contain different information, and the spatial attention mechanism can assist the model in selectively focusing on features from specific locations. By weighting features at different locations, the model can more effectively capture important spatial information and better grasp the differences of objects, thus extracting distinguishing features from objects with different appearances. Therefore, the MSpaARM containing n groups spatial attention residual modules (SpaARM) (as shown in Fig. 1) was designed to establish associations between contexts at different locations and extract discriminative features.

In the SpaARM (as shown in Fig. 3), we adopt a UNet-like architecture [53] with a combination of upsampling and downsampling. After performing upsampling, we elementally add the corresponding feature map from the preceding downsampling stage. This can be seen as the creation of a new branch, lessening the significance of depending on the original attention mask alone of network output. It preserves the good properties of the original features and bypasses the attention mask to propagate to the top level, thus weakening the feature selection capability of the mask branch. Some valuable information may be lost during the attention mask procedure due to upsampling and downsampling. This compensation enables improved transmission of features to deeper layers and enhances network stability, which is crucial for handling complex features in RS data by this new branch.

First, the feature map of the input is defined for MSpaARM, which is derived from the shallow information obtained by TBCNN as follows:

$$F_{\text{in}}^{\text{Spa}} = \text{Concat}(X_H^{\text{out}}, X_L^{\text{out}})$$
$$F^{\text{Spa}} = \text{Conv}_{1\times1}(F_{\text{in}}^{\text{Spa}}) \tag{11}$$

where $F^{\text{Spa}} \in \mathbb{R}^{c\times s\times s}$, c represents the number of multisource data channels, and s is the size of the feature map after TBCNN.

Subsequently, the F^{Spa} is first sliced into n groups of feature maps: F^{Spa_1}, F^{Spa_2},..., F^{Spa_n}, where n is a power of 2 and $n \in \{1, 2, 4, \ldots, c\}$. Second, each group of feature maps goes through the SpaARM to generate the spatial attention maps. Finally, the multiscale spatial feature maps may be generated using these spatial attention maps $F_{\text{out}}^{\text{Spa}} \in \mathbb{R}^{c\times s\times s}$ through (15). The kth group $F^{\text{Spa}_k} \in \mathbb{R}^{c/n\times s\times s}$ is used to demonstrate the SpaARM, where $k \in [1, n]$.

First, the spatial features using 3×3 convolution is extracted to get maps $F_1^{\text{Spa}_k} \in \mathbb{R}^{c/n\times(s-2)\times(s-2)}$, then the deeper spatial features are extracted using 3×3 convolution to obtain the maps $F_2^{\text{Spa}_k} \in \mathbb{R}^{c/n\times(s-4)\times(s-4)}$. The reason for using convolution for downsampling instead of pooling is that pooling loses some details of the feature map. We use the bilinear interpolation to upsample $F_2^{\text{Spa}_k}$ into a feature map $F_3^{\text{Spa}_k} \in \mathbb{R}^{c/n\times(s-2)\times(s-2)}$. The recovery process will inevitably result in the new loss of information, so to compensate for the loss of useful information, $F_1^{\text{Spa}_k}$ is added to $F_3^{\text{Spa}_k}$ to obtain $F_4^{\text{Spa}_k} \in \mathbb{R}^{c/n\times(s-2)\times(s-2)}$, which allows for a better transmission of the features into deeper layers. Similarly, $F_4^{\text{Spa}_k}$ are upsampled to $F_5^{\text{Spa}_k}$ by using bilinear interpolation. Then, a sigmoid function is used to generate the weights W^{Spa}, which are multiplied with the original feature maps, and then, added to the original feature maps to obtain

$$\begin{aligned}
F_1^{\text{Spa}_k} &= \text{Conv}_{3\times3}(F^{\text{Spa}_k}) \\
F_2^{\text{Spa}_k} &= \text{Conv}_{3\times3}(F^{\text{Spa}_k}) \\
F_3^{\text{Spa}_k} &= \text{BiLinear}(F_2^{\text{Spa}_k}) \\
F_4^{\text{Spa}_k} &= F_1^{\text{Spa}_k} + F_3^{\text{Spa}_k} \\
F_5^{\text{Spa}_k} &= \text{BiLinear}(F_4^{\text{Spa}_k}) \\
W^{\text{Spa}} &= \sigma(F_5^{\text{Spa}_k}) \\
F_{\text{out}}^{\text{Spe}_k} &= W^{\text{Spa}} \times F^{\text{Spa}_k} + F^{\text{Spa}_k}
\end{aligned} \tag{12}$$

where $\text{Conv}_{3\times3}(\cdot)$ is the 2-D convolutional layer with kernel size 3×3, and $\text{BiLinear}(\cdot)$ represents the bilinear interpolation.

To construct spatial features F with a size of $c/n \times s \times s$ from n-group spatial maps, the following operation is adopted:

$$F^i = \begin{cases} \Pr(F_{\text{out}}^{\text{Spa}_i}) & i = 1 \\ \Pr(F_{\text{out}}^{\text{Spa}_i} + F_{\text{out}}^{\text{Spa}_{i-1}}) & i = 2 \\ \Pr(F_{\text{out}}^{\text{Spa}_i} + F_{i-1}) & 2 < i \le n. \end{cases} \tag{13}$$

Then, to form the new feature map $F_{\text{out}} \in \mathbb{R}^{c\times s\times s}$, all the features F^i are connected, and a 1×1 convolution is used to enhance the correlation of feature for subsequent summation with the original feature map $F_{\text{in}}^{\text{Spa}}$:

$$F_{\text{out}} = \text{Conv}_{1\times1}(\text{Concat}(F^1, F^2, \ldots, F^n)). \tag{14}$$

Finally, the feature map that contains multiscale spatial information $F_{\text{out}}^{\text{Spa}} \in \mathbb{R}^{c\times s\times s}$ is obtained

$$F_{\text{out}}^{\text{Spa}} = F_{\text{out}} + F_{\text{in}}^{\text{Spa}} \tag{15}$$

E. MCF Module

Multiscale spectral–spatial information can enhance the classification task performance. However, it is important to investigate how to integrate such information, as it is a concern. Inspired by this issue, a simple yet effective MCF (as shown in Fig. 4) module is designed. In the MCF module, we use five parallel convolutional layers to capture different scale information. Combined spectral–spatial feature information $F_{\text{out}}^{\text{Spe}}$ and $F_{\text{out}}^{\text{Spa}}$ as F are fed into model.

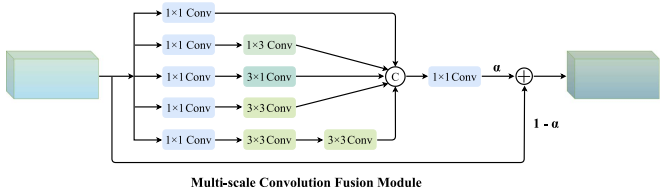


Fig. 4. Illustration of the MCF module.

To be specific, the bottleneck structure is employed in each layers, consisting of 1×1 convolution, to decrease the number of channels in the feature map plus a 3×3 conv-layer. Second, to reduce the number of parameters and introduce deeper nonlinear layers, we replace the 5×5 convolution with two stacked 3×3 convolutions. For the same reason, we use a 1×3 plus a 3×1 convolution to take place of the original 3×3 convolution

$$\begin{aligned}
F_1 &= \text{Conv}_{1\times1}(F) \\
F_2 &= \text{Conv}_{1\times1}(\text{Conv}_{1\times3}(F_1)) \\
F_3 &= \text{Conv}_{1\times1}(\text{Conv}_{3\times1}(F_2)) \\
F_4 &= \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(F_3)) \\
F_5 &= \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{3\times3}(F_4)))
\end{aligned} \tag{16}$$

where $\text{Conv}_{1\times3}(\cdot)$ and $\text{Conv}_{3\times1}(\cdot)$ are the 2-D convolutional layer with kernel sizes 1×3 and 3×1, respectively. Combined the five layers of features and adjust the number of channels with 1×1 convolution as follows:

$$F_{\text{out}} = \text{Conv}_{1\times1}(\text{Concat}(F_1, F_2, \ldots, F_5)). \tag{17}$$

To obtain the output of the whole MCF module, a residual connection with learnable parameter α is then used

$$F_{\text{end}} = \alpha F_{\text{out}} + (1 - \alpha)F. \tag{18}$$

The performance of fused features in the spectral and spatial domains can be improved by successfully integrating the benefits of the two branches using the aforementioned procedures.

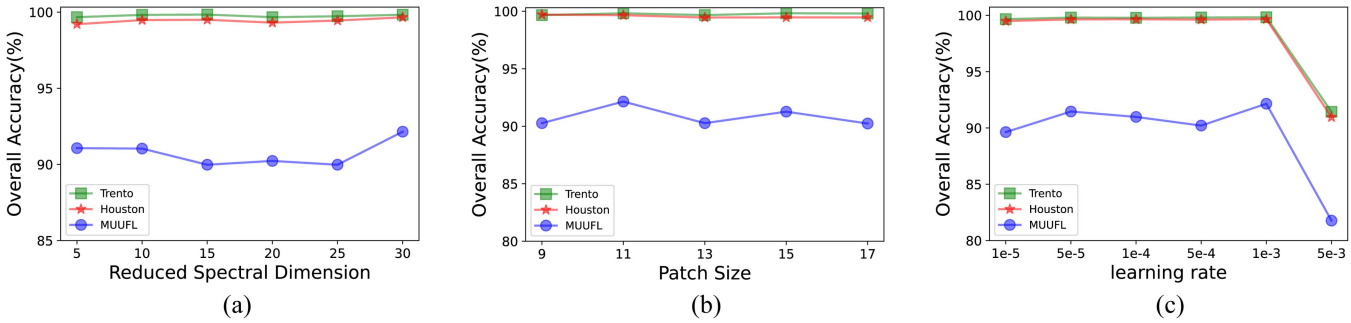


Fig. 5. Impact of different parameters on the OA. (a) Spectral dimension. (b) Patch size. (c) Learning rate.

Algorithm 1: MSSARFNet.

Input: HSI data $X_H \in R^{M\times N\times B}$, LiDAR data $X_L \in R^{M\times N}$, ground-truth data $Y \in R^{M\times N}$.
Output: Classification map **R**.

```
Initialize all weights and bias terms.
Obtain the X_H^pca after PCA transform.
Create all sample patches X_H^P and X_L^P from X_H^pca and X_L, and divide them into a training set and a test set.
for epoch < epochs do
    Perform TBCNN to extract features X_H^out and X_L^out from X_H^P and X_L^P by (1) and (2).
    Perform MSpeARM to extract multiscale spectral features F_out^Spe from X_H^out and X_L^out by (3)–(10).
    Perform MSpaARM to extract multiscale spatial features F_out^Spa from X_H^out and X_L^out by (11)–(15).
    Perform MCF module to aggregate the spectral–spatial information F_out^Spe and F_out^Spa, and obtain the final fusion result F_end by (16)–(18).
    Input the F_end to the classification module, and obtain the predicted value to identify the labels.
end for
Obtain classification map R.
```

F. Classification Module

The F_{end} is fed into a classification module for the final classification. The classification module (as shown in Fig. 1) consists of a maxpool layer, and two FC layers. In the classification module, the spectral–spatial semantic features are extracted using a maxpooling operation, and the two FC layers produce a predicted value. The last linear layer, which is designed to obtain the final labels used in classification, incorporates a softmax function. For each pixel, the category corresponds to the label with the maximum probability. The cross-entropy function is as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{t=1}^{N} q(x_t)\log(p(x_t)) \tag{19}$$

where $q(x_t)$ represents the probability value assigned to class t for the result predicted by the model. $p(x_t)$ represents the expected probability value assigned to class t.

The overall training process of the proposed MSSARFNet is illustrated in Algorithm 1.

IV. Experiments and Analysis

Three HSI-LiDAR RS datasets are used to verify the effectiveness of the proposed network. All experiments are implemented on the PyTorch platform, using an Inter Xeon Silver 4110 2.1-GHz CPU, 128-GB RAM, and an NVIDIA GeForce RTX 2080Ti GPU with 11 GB of RAM. Three commonly used evaluation metrics, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa), are adopted to intuitively quantify the experimental results. In addition, the Adam optimizer was chosen as the initial optimizer to optimize the network. The minibatch size and the number of training epochs were set to 64 and 100, respectively, for the training stage. The general organization is as follows: Section IV-A introduces the datasets used for experiments. The corresponding parameters setting and analyses are performed in Section IV-B. The experimental results of the proposed method, along with a comparison to other methods and an analysis of the results, are presented in Section IV-C. In Section IV-D, ablation experiments are conducted to demonstrate the effectiveness of the various modules within the proposed framework.

In Table I contains a list of three HSI-LiDAR datasets, along with the names of the land cover categories, the number of training samples, and the number of test samples.

A. Data Description

To evaluate the effectiveness of the proposed framework, three HSI-LiDAR datasets are selected for the experiment: Missouri University and University of Florida (MUUFL) Gulfport [54], [55], Houston2013 [1], and Trento [19].

1) *MUUFL Dataset:* Captured in November 2010 at the University of Southern Mississippi, Gulfport Campus in Long Beach, MS, USA. The dataset consists of HSI and LiDAR-based digital surface modeling with a pixel size of 325 × 220, with 72 spectral bands for the hyperspectral data. Due to noise, the first and last eight bands were removed, yielding a total of 64 spectra with a spectral resolution ranging from 375 to 1050 nm and a spatial resolution of 0.54 × 1.0 m. The LiDAR image contains two rasters of elevation data with a resolution of 0.60 × 0.78 m. This dataset consists of 11 different land cover

TABLE I
TRAINING AND TEST SAMPLE NUMBERS IN THE MUUFL DATASET, THE HOUSTON2013 DATASET, AND THE TRENTO DATASET

No.	MUUFL dataset			Houston2013 dataset			Trento dataset		
	Class Name	Training	Test	Class Name	Training	Test	Class Name	Training	Test
C01	Trees	150	23 096	Healthy Grass	198	1 053	Apple Trees	129	3 905
C02	Mostly Grass	150	4 120	Stressed Grass	190	1 064	Buildings	125	2 778
C03	Mixed Ground Surface	150	6 732	Synthetic Grass	192	505	Ground	105	374
C04	Dirt and Sand	150	1 676	Trees	188	1 056	Woods	154	8 969
C05	Road	150	6 537	Soil	186	1 056	Vineyard	184	10 317
C06	Water	150	316	Water	182	143	Roads	122	3 052
C07	Buildings Shadow	150	2 083	Residential	196	1 072			
C08	Buildings	150	6 090	Commercial	191	1 053			
C09	Sidewalk	150	1 235	Road	193	1 059			
C10	Yellow Curb	150	33	Highway	191	1 036			
C11	Cloth Panels	150	119	Railway	181	1 054			
C12				Paking Lot1	192	1 041			
C13				Paking Lot2	184	285			
C14				Tennis Court	181	247			
C15				Running Track	187	473			
-	Total	1 650	52 037	Total	2 832	12 197	Total	819	29 395

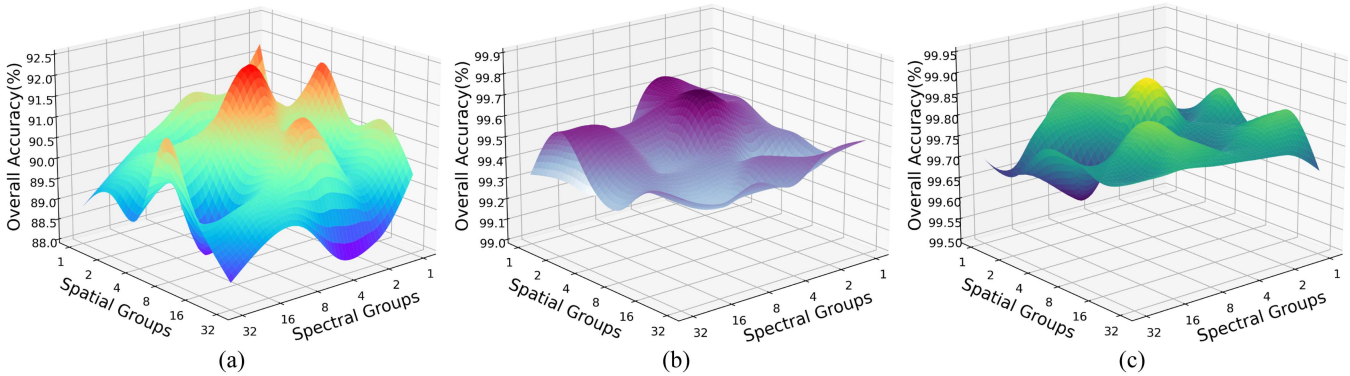


Fig. 6. Classification performance of different numbers of both the spectral and the SpaARM on three HSI-LiDAR datasets. (a) MUUFL dataset. (b) Houston2013 dataset. (c) Trento dataset.

classes. Fig. 7(a)–7(c) visualizes a pseudocolor composite image of the HSI data, a grayscale image of the LiDAR data, and a ground-truth map.

2) *Houston2013 Dataset:* The University of Houston dataset was collected in 2013 as part of the IEEE Geoscience and Remote Sensing Society data fusion competition using the Compact Airborne Spectral Imager. The dataset comprises HSI and LiDAR imagery, covering the University of Houston campus and the surrounding urban area in Houston, TX, USA. Both the HSI and LiDAR images have dimensions of 349 $\times$ 1905 pixels and a spatial resolution of 2.5 m. There are 144 spectral bands in the HSI data, and their spectral resolutions range from 0.38 to 1.05 μm. Only LiDAR images are provided for the same area. The ground truth of the dataset has 15 different land cover classes. Fig. 8(a)–(c) visualizes a pseudocolor composite image of the HSI data, a grayscale image of the LiDAR data, and a ground-truth map.

3) *Trento Dataset:* Captured in a rural area south of Trento, Italy, with 63 bands of HSI in the wavelength range $0.42 - 0.99$ μm and one raster in the LiDAR data providing elevation information. The spectral resolution is 9.2 nm and the spatial resolution is 1 m per pixel. The pixel sizes are all 600 $\times$ 166 and the dataset describes six different land cover classes. Fig. 9(a)–9(c) visualizes a pseudocolor composite image of the HSI data, a grayscale image of the LiDAR data, and a ground-truth map.

B. Parameter Setting and Analysis

Several hyperparameters that may affect the classification performance were set and analyzed, including the spectral dimension b, the patch size p of input data patches, the learning rate, the number of the SpeARM, and SpaARM.

1) *Spectral Dimension:* To evaluate the effect of the spectral dimension, b are chosen for comparison from the set

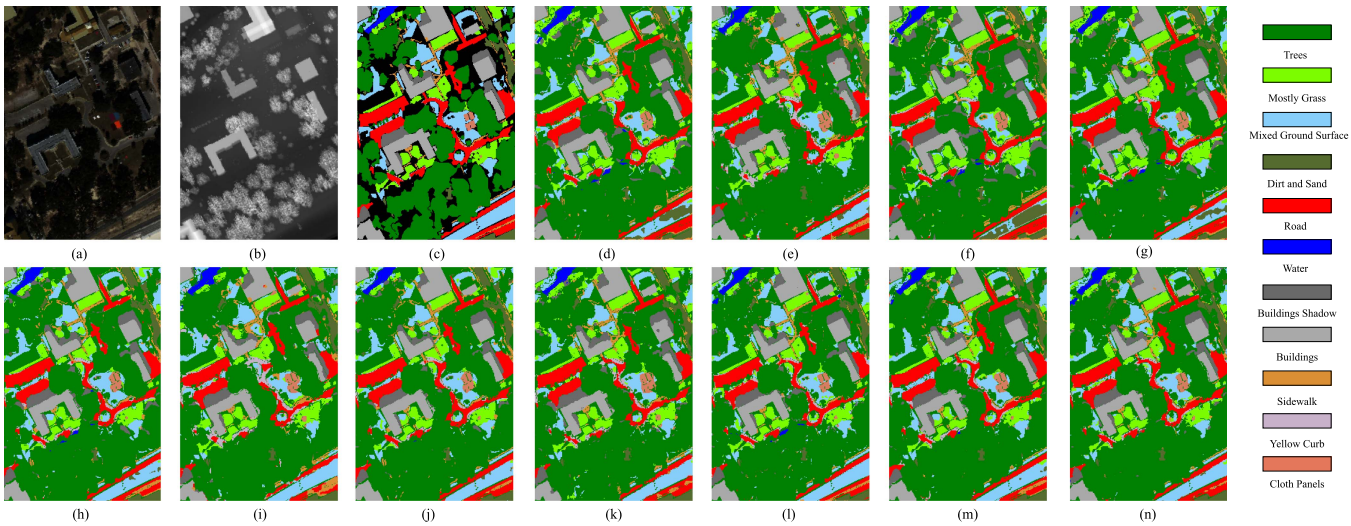


Fig. 7. Classification maps obtained by different methods on the MUUFL dataset. (a) Pseudocolor composite image for HSI. (b) Grayscale image for LiDAR. (c) Ground-truth map. (d) SVM (81.65%). (e) SSFTT (87.56%). (f) S2FL (77.62%). (g) EndNet (82.84%). (h) CoupledCNN (87.22%). (i) MFT (85.10%). (j) FGCN (89.04%). (k) HCT (88.47%). (l) AMSSE-Net (91.41%). (m) MACN (90.58%). (n) Ours (92.15%).

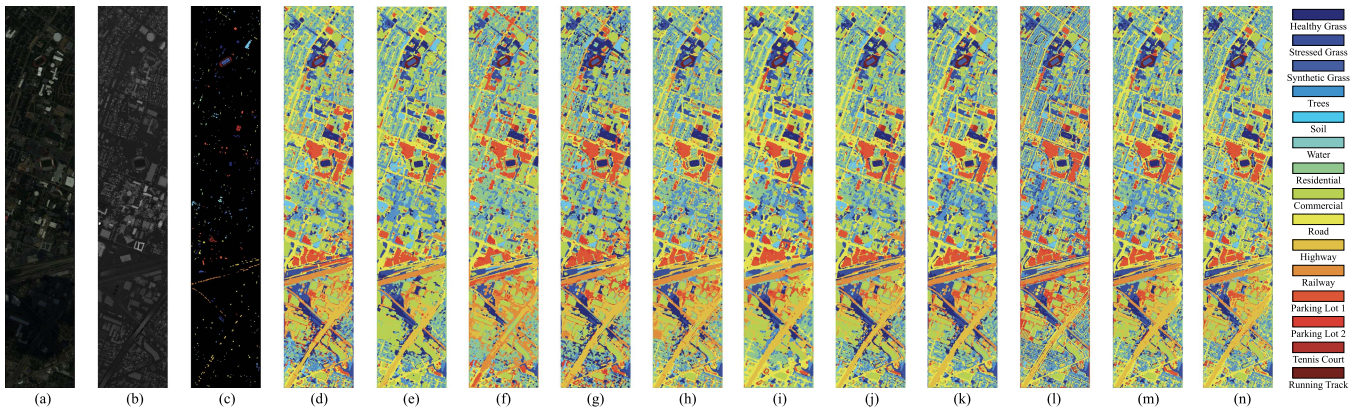


Fig. 8. Classification maps obtained by different methods on the Houston2013 dataset. (a) Pseudocolor composite image for HSI. (b) Grayscale image for LiDAR. (c) Ground-truth map. (d) SVM (93.34%). (e) SSFTT (99.57%). (f) S2FL (81.26%). (g) EndNet (90.18%). (h) CoupledCNN (95.48%). (i) MFT (99.41%). (j) FGCN (98.50%). (k) HCT (99.78%). (l) AMSSE-Net (92.26%). (m) MACN (99.80%). (n) Ours (99.67%).

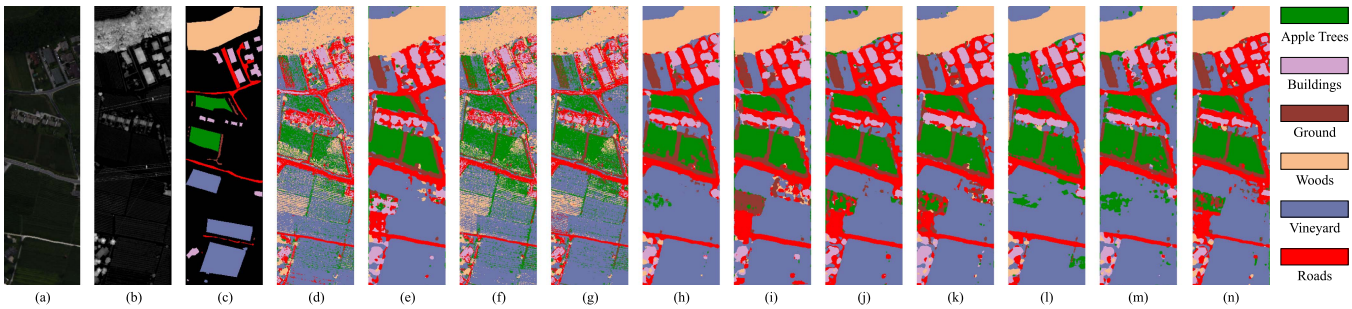


Fig. 9. Classification maps obtained by different methods on the Trento dataset. (a) Pseudocolor composite image for HSI. (b) Grayscale image for LiDAR. (c) Ground-truth map. (d) SVM (82.44%). (e) SSFTT (99.20%). (f) S2FL (85.11%). (g) EndNet (92.92%). (h) CoupledCNN (98.79%). (i) MFT (99.09%). (j) FGCN (99.32%). (k) HCT (99.60%). (l) AMSSE-Net (98.79%)., (m) MACN (99.73%). (n) Ours (99.83%).

TABLE II
COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH DIFFERENT METHODS FOR THE MUUFL DATASET

No.	Class Name	SVM [13]	SSFTT [44]	S2FL [17]	EndNet [26]	CoupledCNN [21]	MFT [46]	FGCN [28]	HCT [31]	AMSSE-Net [29]	MACN [48]	Ours
C01	Trees	83.38	89.56	81.51	84.16	86.01	88.44	90.36	89.45	92.10	91.77	**93.33**
C02	Mostly Grass	84.40	85.90	76.86	83.23	87.26	86.35	88.53	86.02	88.78	85.51	**88.98**
C03	Mixed Ground Surface	68.77	79.19	69.24	70.86	77.66	76.74	82.04	81.44	80.34	82.73	**83.27**
C04	Dirt and Sand	84.93	93.08	78.59	85.70	93.60	96.24	93.73	94.88	96.30	95.51	**97.73**
C05	Road	88.28	84.62	84.21	87.77	89.59	76.38	90.87	84.10	91.01	86.71	**93.05**
C06	Water	93.61	99.68	93.53	94.34	98.08	99.36	100.00	99.64	99.37	100.00	**100.00**
C07	Buildings Shadow	86.42	95.01	84.16	87.74	90.83	90.01	85.76	94.54	94.87	94.52	**95.10**
C08	Buildings	83.25	92.61	76.33	81.73	95.40	88.03	92.08	93.81	94.58	93.23	**97.49**
C09	Sidewalk	75.78	78.87	71.53	75.37	69.94	67.13	78.40	77.84	81.68	81.56	**82.51**
C10	Yellow Curb	98.01	93.94	92.62	97.47	94.26	96.97	96.60	98.21	93.93	97.91	**100.00**
C11	Cloth Panels	97.98	99.16	84.91	97.64	98.32	98.31	100.00	99.78	99.18	100.00	**100.00**
OA (%)		81.65	87.86	77.62	82.84	87.22	85.10	89.04	88.47	91.41	90.58	**92.15**
AA (%)		85.75	90.17	81.15	85.94	89.14	87.64	90.76	90.89	91.64	91.34	**93.69**
Kappa (×100) (%)		77.72	84.20	72.69	78.72	84.52	80.68	86.13	85.19	88.48	87.66	**89.66**

$\{5, 10, 15, 20, 25, 30\}$. The OA performance of the proposed network is shown in Fig. 5(a) for three datasets with different b. We found that the classification performance is the best when $b = 30$.

2) *Patch Size:* Similar to the analysis of spectral dimension, patch size p was selected from the set $\{9, 11, 13, 15, 17\}$ to evaluate its effect. Fig. 5(b) shows the influence of different p on OA. It found that the best OA value is achieved on all three datasets when $p = 11$.

3) *Learning Rate:* In DL models, learning rate is a crucial hyperparameter that controls how fast the objective function can approach its minimum. To evaluate its impact, we choose the learning rate for our experiment from the set $\{1e\text{-}5, 5e\text{-}5, 1e\text{-}4, 5e\text{-}4, 1e\text{-}3, 5e\text{-}3\}$. Learning rate is an important hyperparameter in DL models, as it determines how quickly the objective function can reach its minimum. Fig. 5(c) shows the OA of the proposed network on three datasets with different learning rates. When the learning rate is $1e$-3, the classification performance of all three datasets achieves the best.

4) *Number of the SpeARM and SpaARM*: The number of SpeARM determines the number of subspaces required for the MSpeARM, while the number of SpaARM determines the number of subspaces needed for the MSpaARM. Therefore, it is crucial to refine the extracted spectral and multiscale spatial features using an appropriate number of SpeARM and SpaARM, respectively. In our experiment, the numbers from the set $\{1, 2, 4, 8, 16, 32\}$ were selected to evaluate its effect. Fig. 6 shows the OA of the proposed network on three datasets with different numbers of the SpeARM and the SpaARM. When both number of the SpeARM and the SpaARM is 4, the proposed network achieves the best OA value on three RS datasets.

C. Experimental Results and Analysis

To validate the effectiveness of the proposed MSSARFNet, comparative experiments were conducted using several representative classification methods, including SVM [13], SSFTT [44], S2FL [17], EndNet [26], CoupledCNN [21], MFT [46], FGCN [28], HCT [31], AMSSE-Net [29], and MACN [48]. To ensure a fair comparison, the network parameters of these methods were set the same as described in their respective articles. And the training and test sample sets for the aforementioned methods were selected at random, as shown in Table I.

1) Quantitative Results and Visual Evaluation: The OA, AA, Kappa, and accuracy for each class of the proposed method on the MUUFL, Houston2013, and Trento datasets are presented in the findings as a qualitative categorization of different data in Tables II, III, and IV. The best results are shown in bold. For each of the three datasets, the classification maps of several comparison methods are shown in Figs. 7–9. When compared to the other methods, the maps show that the proposed method yields an excellent classification performance with more distinct boundaries, which is consistent with the numerical results.

For the MUUFL dataset, an OA of 92.15% is obtained, moreover, it is the best in all 11 categories, and the accuracy of $C04$ $DirtandSand$, $C05$ $Road$, and $C08$ $Buildings$ is greatly improved due to the use of the two multiscale attentional residual modules. From Table II, it can be seen that the results of $C05$ $Road$, $C08$ $Buildings$, and $C09$ $Sidewalk$ do not perform very well except for our method, and we can see from the ground truth that the sample distributions of the three categories are mostly adjacent and relatively scattered. The reason for the poor classification performance of these two categories is due to the fact that the categories have the same land cover material, very similar spectral reflectance curves, and similar elevations, resulting in many misclassified samples between the two categories. Contrarily, our method improves 4% on $C05$ $Road$ and $C08$ $Buildings$ and also improves 1% on $C09$ $Sidewalk$ and as evident from the results presented in Fig. 7, the features extracted by the proposed network show good continuity and smoothness within these regions. In addition, the boundaries between different material regions are sharper in our results, closely resembling the ground truth.

For the Houseton2013 dataset, from Table III, our results are not the best in the three commonly used evaluation metrics, which are lower than HCT and MACN, and ranked in the third position. We analyzed that our network failed to extract

TABLE III
COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH DIFFERENT METHODS FOR THE HOUSTON2013 DATASET

No.	Class Name	SVM [13]	SSFTT [44]	S2FL [17]	EndNet [26]	CoupledCNN [21]	MFT [46]	FGCN [28]	HCT [31]	AMSSE-Net [29]	MACN [48]	Ours
C01	Healthy Grass	95.44	100.00	97.82	96.13	93.64	97.82	98.72	99.48	83.10	99.53	99.05
C02	Stressed Grass	97.85	99.87	87.11	94.72	96.86	99.72	98.30	99.84	85.15	99.83	**99.91**
C03	Synthetic Grass	100.00	100.00	99.02	98.73	99.39	100.00	98.87	99.84	83.56	99.82	**100.00**
C04	Trees	94.61	99.71	92.20	91.50	97.37	**100.00**	98.54	99.74	93.37	99.81	99.91
C05	Soil	99.11	100.00	98.50	98.98	99.30	100.00	99.91	100.00	100.00	100.00	**100.00**
C06	Water	96.92	100.00	84.83	93.22	90.25	100.00	100.00	99.99	93.01	100.00	**100.00**
C07	Residential	92.11	**99.81**	83.18	89.25	95.89	98.32	98.15	99.64	95.34	99.47	99.25
C08	Commercial	93.81	99.71	53.25	76.52	92.49	99.62	97.68	99.43	91.74	100.00	**100.00**
C09	Road	88.34	98.39	80.52	86.05	90.40	99.43	97.27	**99.88**	94.24	99.53	99.06
C10	Highway	94.13	99.90	73.11	91.28	96.53	100.00	98.87	100.00	96.20	100.00	**100.00**
C11	Railway	95.79	99.15	64.75	87.60	98.28	100.00	97.58	99.98	96.11	100.00	99.91
C12	Paking Lot1	88.48	99.42	69.24	83.18	91.01	98.27	98.18	99.66	99.62	99.90	99.13
C13	Paking Lot2	81.66	96.49	43.51	84.18	92.04	**100.00**	98.97	99.72	81.40	99.65	99.64
C14	Tennis Court	99.53	100.00	98.95	98.33	99.74	100.00	100.00	100.00	100.00	100.00	**100.00**
C15	Running Track	99.70	100.00	99.18	98.99	99.89	100.00	100.00	100.00	99.00	100.00	**100.00**
OA (%)		93.34	99.57	81.26	90.18	95.48	99.41	98.50	99.78	92.26	**99.80**	99.67
AA (%)		94.50	99.50	81.78	90.28	95.51	99.54	98.74	99.81	92.10	**99.81**	99.72
Kappa (×100) (%)		93.88	99.54	79.66	88.72	95.11	99.36	98.38	99.77	91.59	**99.78**	99.64

TABLE IV
COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH DIFFERENT METHODS FOR THE TRENTO DATASET

No.	Class Name	SVM [13]	SSFTT [44]	S2FL [17]	EndNet [26]	CoupledCNN [21]	MFT [46]	FGCN [28]	HCT [31]	AMSSE-Net [29]	MACN [48]	Ours
C01	Apple Trees	89.49	99.64	80.53	87.50	99.54	99.15	99.32	99.10	99.44	99.64	**99.94**
C02	Buildings	76.47	94.67	79.86	87.36	97.61	98.42	96.88	98.95	99.42	**99.57**	98.81
C03	Ground	97.70	100.00	90.21	98.23	98.34	99.73	96.04	100.00	81.82	99.96	**100.00**
C04	Woods	93.63	100.00	93.04	98.34	100.00	99.92	100.00	100.00	100.00	100.00	**100.00**
C05	Vineyard	73.53	100.00	82.58	93.68	100.00	99.99	99.68	99.99	99.81	100.00	**100.00**
C06	Roads	73.91	97.61	80.32	86.59	97.69	94.07	98.78	98.45	92.46	98.84	**99.50**
OA (%)		82.44	99.20	85.11	92.92	98.79	99.09	99.32	99.60	98.79	99.73	**99.83**
AA (%)		84.12	98.65	84.43	91.78	97.70	98.55	98.45	99.34	95.49	99.67	**99.71**
Kappa (×100) (%)		77.03	98.93	80.21	90.53	98.38	98.78	99.09	99.46	98.38	99.66	**99.77**

complete species information from this dataset compared to the HCT and MACN model, and it also did not fully utilize the extracted information during the fusion process. This is a point that needs to improve in the future. However, our model are superior to HCT and MACN in some individual classes. For classes with similar materials, e.g., $C02 - C04$, grass, and tree, respectively, the proposed method achieves 99.91%, 100%, and 99.91% accuracy. As shown in Fig. 8, they are clearly distinguishable. It is shown that the proposed network has some differentiation between classes composed of similar materials.

Some of the methods suffer from noise due to loss of spatial information or inadequate extraction of spatial–spectral information. For the Trento dataset, in particular, it can be seen in Fig. 9(e)–(g) that it is difficult for SVM to maintain spatial continuity due to the lack of spatial information. The spectral–spatial information is not sufficiently extracted making it difficult for S2FL and EndNet to maintain spectral–spatial smoothing, which results in a lower accuracy of 10% as listed in Table IV. On the contrary, in homogenous regions, our method and other methods produce smoother results, and thus, superior classification performance. For $C06$ $Roads$, the accuracy reaches 99.50% in this category compared to other methods.

The aforementioned three HSI-LiDAR datasets have validated the effectiveness of the proposed model.

2) Robustness to Percentage of Training Samples: As shown in Fig. 10, 2%, 4%, 6%, and 8% labeled samples were randomly selected as training data for the MUUFL and Trento datasets, and 5%, 10%, 15%, and 20% labeled samples were randomly selected as training data for the Houston2013 dataset in order to measure the stability and robustness of the proposed method with different percentages of training samples. Even with a limited number of samples, the proposed method still performs well. Furthermore, for the MUUFL dataset, other methods have lower accuracy when the sample percentage is 2%. Since the accuracy was about 100% even for the small samples, the OA did not show a significant change for the Houston2013 and Trento datasets as the number of samples rose. It is evident from the experiments conducted on the three datasets that the proposed method consistently achieved the best classification results across the entire range of sample sizes.

3) Time Cost Comparison: As shown in Table V, we compared the computation time of our method, including SSFTT, EndNet, CoupledCNN, MFT, FGCN, HCT, AMSSE-Net, and MACN. It is obvious that our method is relatively faster than other methods except EndNet. Therefore, our method can effectively reduce the computation time and improve the classification efficiency. SSFTT, MFT, HCT, and MACN adopt the transformer architecture with more attention layers, which make

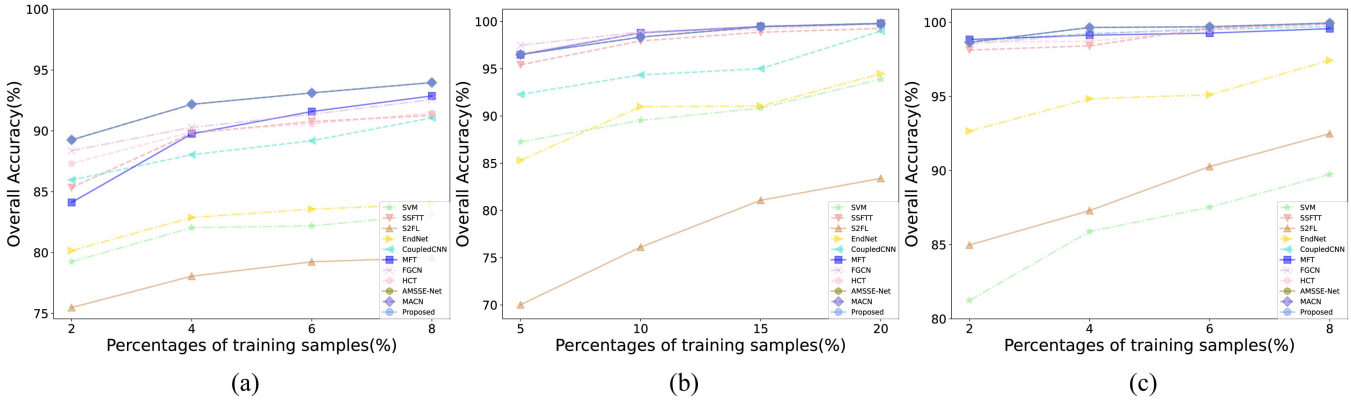


Fig. 10. Classification performance with different training samples percentages. (a) MUUFL dataset. (b) Houston2013 dataset. (c) Trento dataset.

TABLE V
TRAINING TIME IN MINUTES AND TEST TIME IN SECONDS BETWEEN THE CONTRAST METHODS AND PROPOSED METHOD

Methods	GFLOPs(G)	Params(M)	MUUFL		Houston2013		Trento	
			Train.(m)	Test(s)	Train.(m)	Test(s)	Train.(m)	Test(s)
SSFTT	0.45	0.15	7.32	8.13	6.78	4.98	3.12	6.03
EndNet	3.54	0.35	**1.75**	**4.26**	**1.61**	**3.12**	**2.63**	**5.14**
CoupledCNN	0.17	0.11	2.67	7.98	3.16	4.23	2.98	5.97
MFT	0.31	0.22	7.39	8.16	6.85	5.15	3.25	6.18
FGCN	8.77	1.43	4.13	8.02	5.76	4.48	4.46	6.54
HCT	0.51	0.43	7.83	8.27	7.35	5.02	5.44	6.58
AMSSE-Net	2.91	1.31	2.72	4.85	5.35	3.82	3.67	5.79
MACN	0.47	0.33	5.22	8.13	6.06	4.78	4.98	6.32
Ours	9.58	3.06	2.65	4.58	3.18	2.79	2.72	5.56

TABLE VI
ABLATION ANALYSIS OF DIFFERENT MODAL DATA INPUTS

Cases	Indicators	MUUFL	Houston2013	Trento
Only HSI	OA (%)	89.23	99.23	99.41
	AA (%)	91.93	99.31	99.02
	Kappa (×100)(%)	85.95	99.21	99.32
Only LiDAR	OA (%)	69.77	75.36	91.36
	AA (%)	70.61	75.74	86.64
	Kappa (×100)(%)	61.58	70.87	87.21
HSI+LiDAR	OA (%)	**92.15**	**99.67**	**99.83**
	AA (%)	**93.69**	**99.72**	**99.71**
	Kappa (×100) (%)	**89.66**	**99.64**	**99.77**

their training and testing process take a relatively long time, and each iteration consumes a large number of computation cycles.

D. Ablation Analysis

1) Due to the diverse impact of different modalities of input data on the classification performance of the model, three sets of experiments were designed: single LiDAR data input, single HSI data input, joint input of HSI, and LiDAR. The experimental results, as presented in Table VI, demonstrate that the proposed network achieved superior classification results by comparing the OA, AA, and Kappa metrics across the three datasets. This indicates that both HSI and LiDAR data positively contribute to enhancing the classification performance. Furthermore, it confirms that the network effectively utilizes the valuable information from different modalities of RS data.

2) Considering the influence of different components in the network on the classification performance of the model, ablation experiments were conducted to analyze the individual contributions to the classification performance. Specifically, the proposed MSSARFNet in Section III includes components such as spectral feature extraction, spatial feature extraction, spectral–spatial joint extraction, and multiscale feature fusion. The utilization of different modules in the network, as shown in Table VII, including feature extraction and fusion, can have an impact on the overall performance. In Table VII, the MSpeARFN represents a network that uses only spectral information, the MSpaARFN represents a network that uses only spatial information, spectralspatial Network (MSSARN) is a spectral–spatial unification network, and the MSSARFN is with MCF module. Taking the MUUFL dataset as an example, it was observed that both MSpeARFN and MSpaARFN resulted in a decrease in OA by 0.65% and 0.64%, respectively, compared to MSSARN. This implies that while utilizing both spectral and spatial information simultaneously might improve the classification accuracy, relying solely on either one may result in a decline in performance. In contrast to MSSARFN, the MSSARN reduced the OA by 1.09%. Therefore, it is confirmed that multiscale spectral–spatial feature fusion can enhance the classification performance. Quantitative analysis demonstrates that the separability between different classes may be effectively demonstrated via the successful utilization of several processes. The effect of different steps on the three datasets is evident in the T-SNE [56] visualization shown in Fig. 11. It can be observed that the different categories within the datasets become more distinguishable after these steps. This improved distinguishability contributes to achieving better joint classification accuracy.

In conclusion, when analyzing the effect of each component on the three datasets, it is observed that in MSpeARFN, MSpaARFN, and MSSARN, the OA of the classification is lower compared to the MSSARFN. This demonstrates that solely relying on one type of the feature information or not having an effective module to combine the two types of information does not yield better classification results. It further validates the effectiveness of each component within the network.

TABLE VII
ABLATION ANALYSIS OF DIFFERENT COMPONENTS IN MODAL OF OA(%), AA (%), KAPPA×100(%) ON THE THREE DATASETS

Cases	Spectral	Spatial	Spectral–Spatial	MCF	Methods	MUUFL			Houston2013			Trento		
						OA (%)	AA (%)	Kappa×100(%)	OA (%)	AA (%)	Kappa×100(%)	OA (%)	AA (%)	Kappa×100(%)
1	✓	–	–	✓	MSpeARFN	90.41	92.34	87.45	98.21	97.00	98.05	99.63	99.35	99.42
2	–	✓	–	✓	MSpaARFN	90.42	92.36	87.36	98.56	98.52	98.43	99.67	99.48	99.49
3	–	–	✓	–	MSSARN	91.06	92.59	87.78	99.48	99.59	99.43	99.70	99.51	99.61
4	✓	✓	✓	✓	MSSARFN	**92.15**	**93.69**	**89.66**	**99.67**	**99.72**	**99.64**	**99.83**	**99.71**	**99.77**

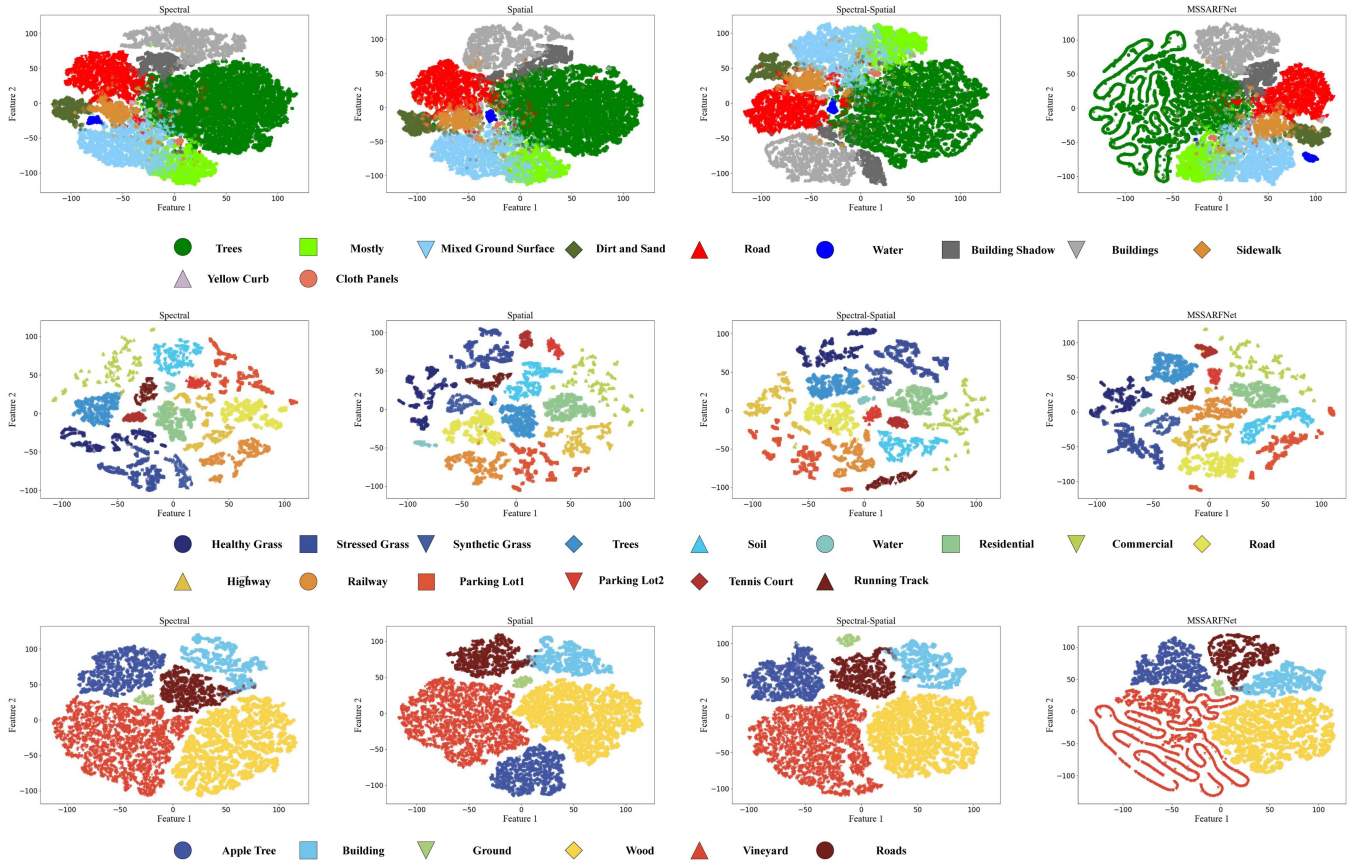


Fig. 11. Feature visualization for different feature extraction strategy in three RS datasets. (Top) MUUFL. (Middle) Houston2013. (Bottom) Trento.

V. CONCLUSION

In this article, an MSSARFNet is proposed for accurate joint classification of HSI and LiDAR data. Combining spectral–spatial information based on a network of attention mechanism, a novel network framework is designed for effective spectral–spatial feature extraction. A effective MCF module is designed to fuse the spectral–spatial features. Three multisource RS datasets are used for experiments, and the findings demonstrate that the method performs excellent classification results compared with current classification methods. In the future, most of the existing classification of multisource data is based on the time-domain aspect, our next step is to consider designing an attention mechanism in the frequency domain, and a joint spectral–spatial attention mechanism in the time domain to form a three-branch attention mechanism to extract multisource data information.

REFERENCES

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

[2] Y. Ding et al., "Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5536716, pp. 1–16, Aug. 2022.

[3] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5513412, pp. 1–12, May 2023.

[4] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[5] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5527812, pp. 1–12, Oct. 2023.

[6] D. Hong et al., "SpectralGPT: Spectral foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, doi: 10.1109/TPAMI.2024.3362475.

[7] S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza, "Morphological convolutional neural networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8689–8702, Jun. 2021.

[8] Y. Ding, X. Zhao, Z. Zhang, W. Cai, N. Yang, and Y. Zhan, "Semi-supervised locality preserving dense graph neural network with ARMA filters and context-aware learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5511812, pp. 1–12, Aug. 2022.

[9] Y. Ding et al., "AF2GNN: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification," *Inf. Sci.*, vol. 602, pp. 201–219, 2022.

[10] Z. Wang and M. Menenti, "Challenges and opportunities in LIDAR remote sensing," *Front. Remote Sens.*, vol. 2, 2021, Art. no. 641723.

[11] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[12] Z. Ye, H. Li, Y. Song, J. A. Benediktsson, and Y. Y. Tang, "Hyperspectral image classification using principal components-based smooth ordering and multiple 1-D interpolation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 1199–1209, Feb. 2017.

[13] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[14] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

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

[16] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[17] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.

[18] 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.

[19] 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.

[20] Q. Feng, D. Zhu, J. Yang, and B. Li, "Multisource hyperspectral and LiDAR data fusion for urban land-use mapping based on a modified two-branch convolutional neural network," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 1, 2019, Art. no. 28.

[21] 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.

[22] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5517010, pp. 1–10, Nov. 2021.

[23] 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, pp. 1–12, 2021.

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

[25] D. Rao, M. De Deuge, N. Nourani-Vatani, S. B. Williams, and O. Pizarro, "Multimodal learning and inference from visual and remotely sensed data," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 24–43, 2017.

[26] 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, no. 5500205, pp. 1–5, Aug. 2020.

[27] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[28] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional Gabor convolutional network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5503818, pp. 1–18, Mar. 2021.

[29] J. Wang, J. Li, Y. Shi, J. Lai, and X. Tan, "AM ^{3} Net: Adaptive mutual-learning-based multimodal data fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5411–5426, Aug. 2022.

[30] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial–spectral enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5531317, Nov. pp. 1–17, 2023.

[31] W. Dong, T. Yang, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Joint contextual representation model-informed interpretable network with dictionary aligning for hyperspectral and LiDAR classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6804–6818, Apr. 2023.

[32] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[33] Y. Zhang, Y. Chen, C. Huang, and M. Gao, "Object detection network based on feature fusion and attention mechanism," *Future Internet*, vol. 11, no. 1, 2019, Art. no. 9.

[34] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9167–9176.

[35] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1307.

[36] Z. Zhang et al., "Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification," *Expert Syst. Appl.*, vol. 217, 2023, Art. no. 119508.

[37] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 582.

[38] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[40] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

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

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

[43] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5518615, pp. 1–15, Nov. 2021.

[44] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5522214, pp. 1–14, Jan. 2022.

[45] Y. Ding et al., "Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 223, Aug. 2023, Art. no. 119858.

[46] 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. 6, Jun. 2023, Art no. 5515620.

[47] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5500716, pp. 1–16, Dec. 2023.

[48] K. Li, D. Wang, X. Wang, G. Liu, Z. Wu, and Q. Wang, "Mixing self-attention and convolution: A unified framework for multi-source remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5523216, pp. 1–20, Aug. 2023.

[49] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExVIT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5514415, pp. 1–15, Jun. 2023.

[50] X. Zhao et al., "Fractional Fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2314–2326, Feb. 2024.

[51] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[52] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[54] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "MUUFL Gulfport hyperspectral and LiDAR airborne data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, 2013.
[55] X. Du and A. Zare, "Technical report: Scene label ground truth map for MUUFL Gulfport data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. 20170417, 2017.
[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Xu Wang received the B.E. degree in computer science and technology from the College of Computer Science and Technology, North China University of Technology, Beijing, China, in 2022. He is currently working toward the M.S. degree in computer science and technology with Xidian University, Xi'an, China .

His research interests include image classification and multisource remote sensing.

Gang Liu was born in Yueyang, Hunan, China. He received the M.S and Ph.D. degrees in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2004, respectively.

From 2005 to 2006, he was an Associate Professor with the Guangdong University of Petrochemical Technology, Guangdong, China. He is currently a Professor with the School of Computer Science and Technology, Xidian University, Xi'an. His main research interests include embedded system, information security, image and video processing, high-speed computer network, and multimedia technology.

Ke Li (Graduate Student Member, IEEE) received the B.E. degree in computer science and technology from the College of Computer Science and Technology, Shandong Jianzhu University, Jinan, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with Xidian University, Xi'an, China.

His research interests include machine learning and multisource remote sensing.

Min Dang (Graduate Student Member, IEEE) received the M.S. degree in computer science and technology in 2022 from Xidian University, Xi'an, China, where she is currently working toward the Ph.D. degree in computer science and technology.

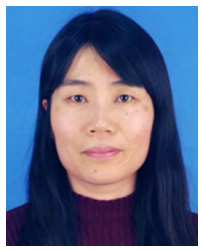
Her current research interests include object detection, action recognition, and behavior detection.

Di Wang (Member, IEEE) received the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2016.

She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. She has authored and coauthored several scientific articles in refereed journals including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and conferences including the SIGIR and International Joint Conferences on Artificial Intelligence. Her research interests include machine learning and multimedia information retrieval.

Zili Wu is a Senior Engineer. He received the M.S. degree in computer system architecture from Xidian University, Xi'an, China, in March 1994, and a Senior Engineer Technical Title from Xidian University, Xi'an, China, in July 1999. He is currently an Associate Professor and master's supervisor with the School of Computer Science and Technology, Xidian University, Xi'an, China.

His current research interests include system architecture, embedded control, and image processing and recognition.



Rong Pan was born in Xingping, Shaanxi, China. She received the B.E. and Ph.D. degrees in computer applied technology in 1997 and 2005, respectively, from Xidian University, Xi'an, China, where she is currently an Associate Professor and master's supervisor with the School of Computer Science and Technology.

Her research interests include image processing and recognition and computer vision.