Spectral-spatial multi-layer perceptron network for hyperspectral image land cover classification

Xiong Tan and Zhixiang Xue

The PLA Strategic Support Force Information Engineering University, Zhengzhou, Henan, China

ABSTRACT
This paper proposes a novel spectral-spatial multi-layer perceptron network for hyperspectral image land cover classification. Current deep learning-based methods have limitations in spectral and spatial feature representation of hyperspectral images, and these shortcomings will severely restrict the hyperspectral image classification performance. The proposed spectral-spatial multi-layer perceptron network exclusively utilizes multi-layer perceptron to represent and classify hyperspectral images. Specifically, the spectral multi-layer perceptron is investigated to model the long-range dependencies along the spectral dimension, because all diagnostic spectral bands contribute to classification performance. Then, we exploit the spatial multi-layer perceptron to extract local spatial features from hyperspectral data, which are also crucial for land cover classification. Furthermore, global spectral characteristics and local spatial features are integrated to perform the hyperspectral image spectral-spatial classification. Three benchmark hyperspectral datasets are employed for comparative classification experiments and ablation study, and experimental results certify the effectiveness and advancement of the proposed model in terms of collaborative classification accuracy.

Introduction

As one of the most important components of earth observation (EO), remote sensing technology can recognize the observed scenes using their specific reflection and emission characteristics without making physical contact with objects. The imaging spectroscopy obtains approximately continuous spectrum from visible to infrared wavelength ranges, thus acquired hyperspectral images (HSIs) have hundreds of diagnostic spectral bands for subsequent information extracting (Ghamisi et al., 2017). As the most vibrant direction in hyperspectral remote sensing community, the HSI classification aims to assign each pixel to one certain category and has been widely applied in, e.g. land surveying, resource management and urban development (Gao et al., 2018).

Certain essential characteristics of hyperspectral images make it very challenging for land-cover classification task. Containing multiple network layers, deep neural networks can extract deep discriminative features from raw data for land cover classification, and many types of deep learning models have been investigated in this field. The convolutional neural networks (CNNs) can extract multiple abstract discriminative features, which have proved to be very effective in HSI classification field and gained more and more attentions in recent years. According to the types of input features utilized for CNNs, convolutional neural networks classification models can be divided into one-dimensional, two-dimensional and three-dimensional patterns (Chen et al., 2016; Lee & Kwon, 2017; Y. Li et al., 2017; Z. Zhong et al., 2017). Due to the unique advantages of convolutional operations in the field of image processing, the CNNs have dominated the HSI classification field in recent years. However, the CNN model is less effective at exploring the spatial relationships among learned instantiation parameters, such as perspective, size, and orientation. To better handle the spectral and spatial features in the spatial domain, capsule network (CapsNet) models using spectral-spatial capsules have also been developed for HSI classification (Arun et al., 2019; Paoletti et al., 2019). Since HSIs contain both rich spatial and spectral information, CNNs and CapsNet have limitations in the spectral sequence feature extraction. For the purpose of modelling the sequence relationships in hyperspectral imagery, the recurrent neural network (RNN) based methods have also been proposed to exploit both spectral and spatial contexts (Mou et al., 2017). However, RNNs face the challenges of computational complexity and modelling long-range dependencies.

In order to solve the scarce availability problem of labelled samples and improve classification performance, the generative adversarial network (GAN) models using both the training and generated samples are designed to conduct the semi-supervised classification of hyperspectral images (Z. He et al., 2017; Lin...
et al., 2018; Xue, 2020). The graph convolutional network (GCN) constructs graph structures using both labelled and unlabelled samples, and then performs deep feature extraction and classification on the graph structures (Hong et al., 2021; Mou et al., 2020). The above two semi-supervised deep learning models can solve the problem of limited number of labelled samples to a certain extent, but they also face difficulties in sample generation and adjacency matrix construction. As an unsupervised feature extraction method, the autoencoder can perform feature extraction and fusion, which relieves the dependence on labelled samples (Patel & Upal, 2022; Sellami et al.). In recent years, the attention mechanism is introduced into land cover classification field to enhance the representation capability of extracted features, which can adaptively recalibrate spectral and spatial features for better classification (Haut et al., 2019; R. Li et al., 2020; Xue et al., 2021). Very recently, inspired by the transformer in the natural language processing (NLP) field, the visual transformer (ViT) based models have been investigated to model the spectral and spatial features for remote sensing image land cover classification (Bazi et al. 2021; He et al., 2021). In addition to above model-based classification models, several machine learning strategies have also been applied in HSI classification, such as transfer learning (He et al., 2020), domain adaptation (Ma et al., 2019), and meta learning (Liu et al., 2019), etc.

Due to the characteristics of hyperspectral image classification, current classification methods mainly focus on solving the bottleneck problems of fusion of spectral and spatial features as well as limited number of labelled samples. It is a fact that whole spectral information in spectral domain and local spatial features in spatial domain simultaneously contribute to remote sensing image interpretation. Nevertheless, current classification methods have limitations in handling long-range dependencies along the spectral dimension and extracting local spatial features in the spatial domain at the same time, and these heterogeneous features are vitally important for the representation of hyperspectral images. And traditional feature extraction methods (e.g. Gabor filters and morphological profiles) have limitations in extracting deep discriminative features. What is more, spectral and spatial collaborative classification will significantly improve the classification performance, because heterogeneous features have rich and complementary information for land cover classification.

Thus, we propose a novel spectral-spatial multi-layer perceptron network (SSMLP) for hyperspectral image land cover classification using multi-layer perceptron (MLP) as backbone network. Different existing hyperspectral image classification approaches, our proposed method employs multi-layer perceptron network to extract global spectral features and local spatial characteristics, constructing an end-to-end supervised learning architecture, which can learn discriminative spectral and spatial heterogeneous features for land cover classification. The main contributions of this paper can be summarized as follows:

The first approach to utilize multi-layer perceptron, applied to image patches and across patches, for extracting spectral and spatial features, which can represent the global spectral and local spatial information of hyperspectral images effectively.

To better utilize the spectral and spatial features for collaborative classification, we propose to employ different sizes of image blocks to extract spectral and spatial features respectively, and these heterogeneous features are fused for joint classification. The effectiveness and advancement of proposed method have been verified on two benchmark hyperspectral data set.

The remainder of this article is organized as follows. The details of the proposed model are introduced in Section 2. The experimental results are reported and analysed in Section 3. And Section 4 gives the conclusion of this article.

Proposed method

Network Architecture of SSMLP

Due to the fact that hyperspectral image contains rich spatial and spectral information simultaneously, we investigate the double-branch multi-layer perceptron structure to extract spectral and spatial features separately. The framework of proposed SSMLP for hyperspectral image land cover classification is illustrated in Figure 1.

In the network structure, the input of the spectral multi-layer perceptron branch is hyperspectral image blocks, and a series of spectral multi-layer perceptron layers are utilized to extract global spectral characteristics. For the spatial multi-layer perceptron branch, we first conduct the dimension reduction and spatial feature extraction operation, and a number of spatial multi-layer perceptron layers are employed to learn spatial information. After the heterogeneous spectral and spatial features extraction, the global average pooling operations are used in the end of two branches. And the spectral and spatial features are concatenated for final joint land cover classification using the final multi-layer perceptron layer.

Spectral multi-layer perceptron

Since all diagnostic spectral bands contribute to the classification performance, we propose the spectral multi-layer perceptron to extract spectral features, which is shown in Figure 2.

Because spectral bands in hyperspectral images have natural sequence structure, we expand the hyperspectral image patches $\mathbf{R}^{H \times W \times C}$ along the spatial dimension as
input embeddings. This results in a two-dimensional input table $X \in \mathbb{R}^{S \times C}$, where the number of tokens is $S=HW$, $H$ and $W$ represent the length and width of the image, respectively, and the token dimension is $C$. The spectral multi-layer perceptron consists of two MLP blocks, namely the token-mixing MLP and the channel-mixing MLP blocks. In the token-mixing MLP block, input table $X$ is first applied the transpose operation $T$, then it acts on columns of $X$, denoted as $1MLP_1$, maps $\mathbb{R}^{S} \rightarrow \mathbb{R}^{S}$, and is shared across all columns. After the token-mixing MLP operation, and another transpose operation is applied on the $X$. Then, the channel-mixing MLP acts on rows of $X$, denoted as $1MLP_2$, maps $\mathbb{R}^{C} \rightarrow \mathbb{R}^{C}$, and is shared across all rows. There are two fully connected layers and the Gaussian Error Linear Unit (GELU) nonlinear activation function that perform a nonlinear operation on each row of the input data in each MLP block. The GELU nonlinear activation function can be formulated as follows,

$$GELU(x) = x\Phi(x) = 0.5x\left(1 + \tanh\left(\sqrt{2/\pi}(x + 0.044715x^3)\right)\right)$$  \hspace{1cm} (1)

And the layer normalization is also employed in the token-mixing MLP and the channel-mixing MLP blocks, which can be formulated as follows,

$$\bar{x}' = \frac{x' - \mu'}{\sqrt{(\sigma')^2 + \epsilon}}$$  \hspace{1cm} (2)

in which $1x'$, $\mu'$ and $\sigma'$ denotes the feature map, mean and variance of the $l$th layer. The structure of these two MLP blocks can be formulated as follows,

$$U_{s,i} = X_{s,i} + W_2\sigma\left(W_1\text{LayerNorm}(X)_{s,i}\right), i = 1 \cdots D_C$$

$$Y_{j,s} = U_{j,s} + W_4\sigma\left(W_3\text{LayerNorm}(U)_{j,s}\right), j = 1 \cdots D_S$$  \hspace{1cm} (3)

**Spatial multi-layer perceptron**

In the spatial dimension, local spatial features have a greater effect on classification performance, thus we exploit the spatial multi-layer perceptron to extract spatial features in hyperspectral images. The structure of the spatial multi-layer perceptron is illustrated in Figure 3.

We first utilize the invariant attribute profiles (IAPs) to perform the spatial feature extraction and dimensionality reduction, then the image blocks are cut into multiple image patches (Hong et al., 2020). The spatial multi-layer perceptron takes as input a sequence of $S$ non-overlapping image patches, each one projected to a desired hidden dimension $C$. This also results in a two-dimensional input table,
$X \in \mathbb{R}^{S \times C}$. If the original input hyperspectral image block has size $(H, W)$, and each patch has size $(P, P)$, then the number of patches is $S = HW/P^2$, and all patches are linearly projected with the same projection matrix. After the tokens embedding operation, the token-mixing $1MLP_1$ and channel-mixing $1MLP_2$ blocks are conducted to extract local spatial features.

As mentioned above, the same channel-mixing MLP (token-mixing MLP) is applied to every row (column) of $X$. The operation of channel-mixing MLPs provides positional invariance, a prominent feature of convolutions, which is very important for extracting local spatial features. The structural expressions of the two MLP blocks are consistent with formula (3). Furthermore, the skip-connections and layer normalization are also employed in the spatial multi-layer perceptron structure for better training.

**Invariant attribute profiles**

The invariant attribute profiles extract invariant features in both spatial and spectral domains, which consist of spatial invariant features (SIFs) and frequency invariant features (FIFs).

The k-means algorithm is first employed to group the HSI into several groups, then the horizontal and vertical gradients for each group are computed by means of the maximum magnitude response. The polarized Fourier features are extracted by gradients. In the spatial domain, the isotropic spatial filters are employed to get the robust convolutional features (RCFs), then the simple linear iterative clustering (SLIC) is used to generative the SIFs. In the frequency domain, the region-based features are constructed by Fourier convolution kernels, which include order fitting and magnitude maps. After that, the spatial invariant features and frequency invariant features are stacked as the final invariant attribute profiles.

**Experiments and discussion**

**Dataset description**

Our experimental results are conducted on three benchmark HSI datasets including the WHU-Hi-LongKou, Pavia University, and the Houston 2013 hyperspectral datasets.

The WHU-Hi-LongKou hyperspectral dataset was captured over Longkou Town, Hubei province, China using the Headwall Nano-Hyperspectral imaging sensor (ZY. Zhong et al., 2020). There are 270 bands in the wavelength range of 400 to 1000 nm. The spatial size of this imagery is $550 \times 400$ pixels, and the spatial resolution is about 0.463 m. The cover area of this data set is a typical agricultural scene, which contains nine different land-cover classes. And the detailed ground-truth classes and the corresponding number of samples are shown in Table 1.

The Pavia University hyperspectral dataset was acquired by the ROSIS sensor over the Pavia University. There are 103 spectral bands in this dataset. The spatial size of this imagery is $610 \times 340$ pixels, and the geometric resolution is 1.3 m. There are nine different categories within the coverage of this dataset. And the detailed ground-truth classes and the corresponding number of samples are shown in Table 2.

| Table 1. Land-cover classes and samples of the WHU-Hi-LongKou dataset. |
|-----------------------------|----------|------|------|
| NO. | Class name | Train | Test | Percent |
| C1  | Corn       | 200   | 34,311 | 0.58%  |
| C2  | Cotton     | 200   | 8174  | 2.39%  |
| C3  | Sesame     | 2831  | 200   | 6.60%  |
| C4  | Broad-leaf soybean | 200 | 63,012 | 0.32%  |
| C5  | Narrow-leaf soybean | 200 | 3951  | 4.82%  |
| C6  | Rice       | 11,654 | 200   | 1.69%  |
| C7  | Water      | 66,856 | 200   | 0.30%  |
| C8  | Roads and houses | 6924 | 200   | 4.21%  |
| C9  | Mixed weed | 5029  | 200   | 3.82%  |
|     | Total      | 1800  | 202,742 | 0.89% |

| Table 2. Land-cover classes and samples of the Pavia university dataset. |
|-----------------------------|------|------|------|
| NO. | Class name | Train | Test | Percent |
| C1  | Asphalt    | 100   | 6531 | 1.53%  |
| C2  | Meadows    | 100   | 18,549 | 0.54% |
| C3  | Gravel     | 100   | 1999 | 5.00%  |
| C4  | Trees      | 100   | 2964 | 3.37%  |
| C5  | Metal sheets | 100 | 1245  | 8.03%  |
| C6  | Bare soil | 100   | 4929 | 2.03%  |
| C7  | Bitumen    | 100   | 1230 | 8.13%  |
| C8  | Bricks     | 100   | 3582 | 2.79%  |
| C9  | Shadow     | 847   | 100  | 11.81% |
|     | Total      | 900   | 41,876 | 2.15% |
Table 3. Land-cover classes and samples of the Houston 2013 dataset.

| NO. | Class name      | Train | Test | Percent |
|-----|-----------------|-------|------|---------|
| C1  | Healthy grass   | 198   | 1053 | 15.83%  |
| C2  | Stressed grass  | 190   | 1064 | 15.15%  |
| C3  | Synthetic grass | 192   | 505  | 27.55%  |
| C4  | Trees           | 188   | 1056 | 15.11%  |
| C5  | Soil            | 186   | 1056 | 14.98%  |
| C6  | Water           | 182   | 143  | 56.00%  |
| C7  | Residential     | 196   | 1072 | 12.46%  |
| C8  | Commercial      | 191   | 1053 | 15.35%  |
| C9  | Road            | 193   | 1059 | 15.42%  |
| C10 | Highway         | 191   | 1036 | 15.57%  |
| C11 | Railway         | 181   | 1054 | 14.66%  |
| C12 | Parking lot 1   | 192   | 1041 | 15.57%  |
| C13 | Parking lot 2   | 184   | 285  | 39.23%  |
| C14 | Tennis court    | 181   | 247  | 42.29%  |
| C15 | Running track   | 187   | 473  | 28.33%  |
|     | Total           | 2832  | 12,197 | 18.84% |

The Houston 2013 dataset collected by the NSF-funded Center for Airborne Laser Mapping (NCALM) covers the Houston university campus and surroundings (Debes et al., 2014). There are 144 spectral bands covering the wavelength range of 380 to 1050 nm in the hyperspectral image. This dataset has the spatial size of 349 × 1905 pixels, and the corresponding spatial resolution is 2.5 m. This imagery covers typical urban land-cover classes, and there are 15 distinguishable species within the image coverage, and detailed sample information is shown in Table 3.

Experimental setting

In order to evaluate the classification performance of the proposed SSMLP method, we carried out the comparative experiments with several state-of-the-art methods. Compared methods include spectral-based method support vector machine (SVM; Melgani & Bruzzone, 2004), contextual CNN (CDCNN; Lee & Kwon, 2017), spectral-spatial residual network (SSRN; Z. Zhong et al., 2017), double-branch dual-attention mechanism network (DBDA; R. Li et al., 2020), spectral-spatial unified network (SSUN), and hierarchical residual network with attention mechanism (HResNet; Xue et al., 2021).

SVM is a traditional classification method with RBF kernel, which employs the spectral features of HSI for classification. It is mainly based on the kernel theory and can solve the small samples classification problem to a certain degree. The optimal regularization parameter and kernel parameter are selected by the cross-validation strategy.

CNCNN constructs the network with 2D CNN and residual learning structure, which is composed of the multiscale filter band and two residual blocks. It can build very deep networks by the residual blocks.

Figure 4. Box plot of overall accuracies with different learning rates on three hyperspectral datasets: (a) WHU-Hi-LongKou dataset. (b) pavia university dataset. (c) Houston 2013 dataset.

Figure 5. Box plot of overall accuracies with different model depths on three hyperspectral datasets: (a) WHU-Hi-LongKou dataset. (b) pavia university dataset. (c) Houston 2013 dataset.
SSRN extracts spectral and spatial features using 3D CNN, and residual learning strategy is also employed in the model. This method employs different convolution kernels to sequentially extract spatial and spectral features for classification.

DBDA is based on the attention mechanism, which contains spectral and spatial dense blocks and corresponding attention blocks. The attention blocks are used to enhance the discriminative capability of extracted features.

SSUN contains a grouping-based long short-term memory (LSTM) model and multiscale convolutional neural networks to extract spectral and spatial features, and these features and fused for classification.

HResNet is a double-branch network architecture, in which multiscale spectral and spatial features are extracted, and corresponding spectral as well as spatial attention blocks are employed to enhance the discriminative ability of extracted multiscale features.
In SSMLP, the first three invariant attribute profiles are selected as input for spatial branch, and the dimension of tokens in two branches is set as 256. The widely used Adam optimizer is adopted to optimize the proposed model. The number of epochs in the training process is set to be 200, and batch size 64. To increase the credibility of experimental results, we have performed ten trials for each experiment by randomly selecting the labelled samples for training. Aiming to maintaining the balance between spectral and spatial features as much as possible, we regularize the inputs of spectral and spatial branches.

**Parameter sensitivity analysis**

**Learning rate**

The learning rate is an important parameter which controls the step of gradient descent in the SSMLP model. We set the candidate set of this parameter in our experiments as \{0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001\}, and we test the influence of this parameter on classification performance. The overall accuracies with different learning rates on two hyperspectral data sets are reported in Figure 4. From classification experiments using different learning rates, we can see that smaller
learning rate has higher classification accuracy, and the reasonable learning rates are set as 0.00002, 0.00001, and 0.00001, respectively.

Depth of model
Since our SSMLP network mainly consists of MLP layers, the number of MLP layers directly determines the depth of the model. And the model depth affects the feature representation ability, which further influences the classification performance. For the purpose of maintaining the balance between spectral and spatial features as much as possible, the numbers of MLP layers in two branches are consistent. The candidate set of model depth is set as \( \{4, 6, 8, 10, 12, 14\} \), and classification accuracies with different model depths are reported in Figure 5. From this figure, we can see that as the depth of model increases, the overall accuracy first increases and then decreases, which are

![Figure 9. Classification maps of different classification approaches on the Houston 2013 hyperspectral dataset: (a) False colour composite image; (b) ground-truth map; (c) SVM; (d) CDCNN; (e) SSRN; (f) DBDA; (g) SSUN; (h) HResNet; (i) spectral MLP; (j) spatial MLP; (k) SSMLP.](image)

| Table 4. Comparison on the number of parameters and FLOPs (M). |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Datasets          | Complexity       | CDCNN            | SSRN             | DBDA             | SSUN             | HResNet          | Spectral MLP     | Spatial MLP      | SSMLP            |
| WHU-Hi-LongKou    | Parameters       | 1.38             | 0.47             | 0.51             | 0.37             | 0.44             | 1.31             | 1.30             | 2.61             |
|                   | FLOPs            | 4.70             | 98.84            | 44.26            | 76.12            | 24.98            | 18.30            | 18.15            | 36.46            |
| Pavia University  | Parameters       | 0.63             | 0.22             | 0.20             | 0.15             | 0.17             | 1.33             | 1.30             | 2.63             |
|                   | FLOPs            | 1.90             | 48.60            | 20.93            | 30.92            | 13.27            | 40.93            | 18.16            | 59.08            |
| Houston 2013      | Parameters       | 0.81             | 0.28             | 0.28             | 0.23             | 0.24             | 1.73             | 1.73             | 3.46             |
|                   | FLOPs            | 2.59             | 67.94            | 23.34            | 46.71            | 13.23            | 24.26            | 24.20            | 48.45            |

| Table 5. The OA, AA, Kappa coefficient and class-specific accuracy of different approaches for WHU-Hi-LongKou dataset. |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| NO.               | SVM              | CDCNN            | SSRN             | DBDA             | SSUN             | HResNet          | Spectral MLP     | Spatial MLP      | SSMLP            |
| C1                | 98.03            | 99.51            | 99.73            | 99.66            | 99.12            | 99.76            | 99.36            | 98.34            | 99.56            |
| C2                | 74.05            | 89.74            | 93.53            | 94.39            | 99.09            | 98.69            | 94.31            | 95.73            | 97.95            |
| C3                | 52.08            | 78.47            | 97.33            | 92.58            | 97.76            | 95.47            | 89.49            | 91.01            | 97.91            |
| C4                | 98.97            | 99.36            | 99.64            | 99.82            | 97.10            | 99.92            | 99.73            | 99.06            | 99.86            |
| C5                | 55.74            | 71.34            | 86.30            | 88.70            | 96.63            | 82.80            | 96.14            | 93.57            | 97.36            |
| C6                | 97.99            | 98.93            | 99.31            | 99.75            | 99.31            | 99.67            | 98.98            | 96.19            | 99.33            |
| C7                | 99.99            | 99.97            | 99.97            | 99.99            | 98.46            | 99.98            | 99.97            | 99.95            | 99.98            |
| C8                | 89.84            | 89.85            | 95.89            | 97.71            | 98.07            | 96.69            | 93.97            | 95.85            | 97.28            |
| C9                | 80.74            | 83.70            | 93.14            | 88.46            | 99.13            | 93.48            | 95.39            | 95.35            | 98.94            |
| OA                | 94.15 ± 0.91     | 97.15 ± 0.68     | 98.81 ± 0.24     | 98.85 ± 0.17     | 98.75 ± 0.16     | 98.58 ± 0.21     | 97.91 ± 0.20     | 97.10 ± 0.33     | 99.12 ± 0.12     |
| AA                | 83.02 ± 0.83     | 90.10 ± 0.79     | 96.09 ± 0.33     | 95.67 ± 0.18     | 98.52 ± 0.13     | 95.16 ± 0.36     | 96.42 ± 0.32     | 96.12 ± 0.39     | 98.69 ± 0.16     |
| Kappa             | 92.42 ± 0.75     | 96.29 ± 0.88     | 98.44 ± 0.28     | 98.50 ± 0.23     | 98.25 ± 0.14     | 98.14 ± 0.23     | 97.28 ± 0.25     | 96.21 ± 0.36     | 98.70 ± 0.17     |
caused by the underfitting and overfitting phenomena when the model depth is too small or large, respectively. Moreover, we set the reasonable model depth for two hyperspectral data sets as 6, 6, and 8, respectively.

**Labelled samples**
The number of labelled samples for training generally has great influence on the HSI classification performance. To evaluate the impact of the number of labelled samples on classification performance, we randomly choose different numbers of samples per class for training, and the overall accuracies for three benchmark datasets are shown in [Figure 6](#). From this figure, we can observe that the classification accuracy could quickly reach a relatively stable level with the increase of training samples.

**Model complexity**
Model complexity is an important indicator the model, and we conduct a quantitative analysis on the model complexity. Time complexity and space complexity evaluate the model from the perspective of time and space, respectively. The floating-point operations (FLOPs) and the number of parameters are employed to measure the time complexity and space complexity, respectively. And the detailed number of model parameters and FLOPs are reported in [Table 4](#). From this table, we can see that deeper models generally have more parameters and FLOPs. Furthermore, our proposed method has more parameters and FLOPs, but more complex models generally have higher classification accuracies.

**Experiments performance**
The mean and standard deviation of OAs, AAs, kappa coefficients (kappa) and classification accuracies of all categories using different classification methods for WHU-Hi-LongKou, Pavia University, and Houston 2013 datasets are listed in [Tables 5, 6, 7](#), respectively. From these tables, we can see that deeper models have higher classification accuracies than models with fewer layers. Because deeper methods extract more discriminative features for classification. And the spectral-spatial classification methods (e.g. SSRN, DBDA and SSMLP) have better classification performance than spectral or spatial models (e.g. SVM, CDCNN, spectral MLP and spatial MLP), which also prove that our proposed method is an effective spectral-spatial classification model. Among all these classification methods,
our proposed SSMLP has the highest classification accuracy in terms of main classification coefficients, which means that the proposed MLP-based method has strong feature representation capability and classification ability for hyperspectral images.

The classification maps of different methods are also employed to evaluate classification performance from visual perspective. The resulting classification maps with enlarged views of three hyperspectral datasets using different methods are shown in Figure 7, 8, 9, respectively. The ground-truth maps are also displayed together with classification maps to enhance the visual comparison among different methods, and each colour corresponds to a specific land cover category. When comparing classification maps obtained by different methods, we can see that the thematic map produced by the SSMLP has the least classification noise than other algorithms. This is because that our proposed MLP architecture can extract more discriminative spectral and spatial features from hyperspectral images and effectively fuse these heterogeneous features to obtain better classification results.

**Conclusion**

The novel spectral-spatial multi-layer perceptron architecture for hyperspectral image land cover classification has been proposed, which relies on two main aspects, the first one is that we exploit multi-layer perceptron to extract spectral and spatial features in hyperspectral images, the long-range dependencies along spectral dimension and local spatial features are extracted by the spectral multi-layer perceptron and spatial multi-layer perceptron networks, respectively. The second one is that the proposed method can effectively fuse heterogeneous spectral and spatial features for joint classification to further improve the classification performance. Therefore, the spectral-spatial multi-layer perceptron network can extract more discriminative characteristics and effectively fuse heterogeneous features for land-cover classification. The performance of SSMLP has been tested on two benchmark hyperspectral images, and compared with the state-of-the-art traditional and deep learning classification methods. The overall accuracies of SSMLP for three groups of datasets can achieve 99.12%, 98.74%, and 99.49% respectively, which have confirmed the effectiveness and advancement of proposed method. The combination of spatial and spectral features effectively improves the land cover classification performance. In future research, we will study how to extract multimodal features from multi-source remote sensing data for collaborative classification. At the same time, we will also study how to employ self-supervised learning methods to learn feature representations from a large amount of unlabelled data.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**

Arun, P. V., Buhdiraju, K. M., & Porwal, A. (2019). Capsule-net-based spatial-spectral classifier for hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(6), 1849–1865. [https://doi.org/10.1109/JSTARS.2019.2913097](https://doi.org/10.1109/JSTARS.2019.2913097)

Bazi, Y., Bashmal, L., Rahhal, M., Dayil, R., & Ajlan, N. (2021). Vision transformers for remote sensing image classification. *Remote Sensing, 13*(3), 516. [https://doi.org/10.3390/rs13030516](https://doi.org/10.3390/rs13030516)

Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 6232–6251. [https://doi.org/10.1109/TGRS.2016.2584107](https://doi.org/10.1109/TGRS.2016.2584107)

Debes, C., Merentitis, A., Heremans, R., Hahn, J., Frangiadakis, N., van Kasteren, T., Liao, W., Bellens, R., Pizurica, A., Gautama, S., Philips, W., Prasad, S., Du, Q., & Pacifici, F. (2014). Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 2405–2418. [https://doi.org/10.1109/JSTARS.2014.2305441](https://doi.org/10.1109/JSTARS.2014.2305441)

Gao, H., Yang, Y., Li, C., Zhou, H., & Qu, X. (2018). Joint alternate small convolution and feature reuse for hyperspectral image classification. *ISPRS International Journal of Geo-Information, 7*(9), 349. [https://doi.org/10.3390/ijgi7090349](https://doi.org/10.3390/ijgi7090349)

Ghamisi, P., Yokoya, N., Li, J., Liao, W., Liu, S., Plaza, J., Rasti, B., & Plaza, A. (2017). Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 37–78. [https://doi.org/10.1109/MGRS.2017.2762087](https://doi.org/10.1109/MGRS.2017.2762087)

Haut, J., Paolletti, M. E., Plaza, J., Plaza, A., & Li, J. (2019). Visual attention-driven hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 57*(10), 8065–8080. [https://doi.org/10.1109/TGRS.2019.2918080](https://doi.org/10.1109/TGRS.2019.2918080)

He, Z., Liu, H., Wang, Y., & Hu, J. (2017). Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote Sensing, 9*(10), 1042. [https://doi.org/10.3390/rs9101042](https://doi.org/10.3390/rs9101042)

He, X., Chen, Y., & Ghamisi, P. (2020). Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing, 58*(5), 3246–3263. [https://doi.org/10.1109/TGRS.2019.2951445](https://doi.org/10.1109/TGRS.2019.2951445)

He, X., Chen, Y., & Lin, Z. (2021). Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing, 13*(3), 498. [https://doi.org/10.3390/rs13030498](https://doi.org/10.3390/rs13030498)

Hong, D., Wu, X., Ghamisi, P., Chanussot, J., Yokoya, N., & Zhu, X. X. (2020). Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 58*(6), 3791–3808. [https://doi.org/10.1109/TGRS.2019.2957251](https://doi.org/10.1109/TGRS.2019.2957251)
Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., & Chanussot, J. (2021). Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 5966–5978. https://doi.org/10.1109/TGRS.2020.3015157

Lee, H., & Kwon, H. (2017). Going deeper with contextual CNN for hyperspectral image classification. *IEEE Transactions on Image Processing*, 26(10), 4843–4855. https://doi.org/10.1109/TIP.2017.2725580

Li, Y., Zhang, H., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing*, 9(1), 67. https://doi.org/10.3390/rs9010067

Li, R., Zheng, S., Duan, C., Yang, Y., & Wang, X. (2020). Classification of hyperspectral image based on double-branch dual-attention mechanism network. *Remote Sensing*, 12(3):582. doi: https://doi.org/10.3390/rs12030582

Lin, Z., Chen, Y., Ghamisi, P., & Benediktsson, J. (2018). Generative adversarial networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9), 5046–5063. https://doi.org/10.1109/TGRS.2018.2805286

Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G., Wang, R. et al (2019). Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4), 2290–2304. https://doi.org/10.1109/TGRS.2018.2872830

Ma, X., Mou, X., Wang, J., Liu, X., Wang, H., & Yin, B. (2019). Cross-data set hyperspectral image classification based on deep domain adaptation. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12), 10164–10174. https://doi.org/10.1109/TGRS.2019.2931730

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), 1778–1790. https://doi.org/10.1109/TGRS.2004.831865

Mou, L., Ghamisi, P., & Zhu, X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3639–3655. https://doi.org/10.1109/TGRS.2016.2636241

Mou, L., Lu, X., Li, X., & Zhu, X. X. (2020). Nonlocal graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12), 8246–8257. https://doi.org/10.1109/TGRS.2020.2973363

Paoletti, M., Haut, J., Ruben Fernandez-Beltran, J. P., Pla, F., Pla, F., Pla, F., & Pla, F. (2019). Capsule networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4), 2145–2160. https://doi.org/10.1109/TGRS.2018.2871782

Patel, H., & Upla, K. P. (2022). A shallow network for hyperspectral image classification using an autoencoder with convolutional neural network. *Multimedia Tools and Applications*, 81(1), 695–714. https://doi.org/10.1007/s11042-021-11422-w

Sellami, A., & Tabbone, S. (2022). Deep neural networks-based relevant latent representation learning for hyperspectral image classification. *Pattern Recognition*, 121, 108224. https://doi.org/10.1016/j.patcog.2021.108224

Xue, Z. (2020). A general generative adversarial capsule network for hyperspectral image spectral-spatial classification. *Remote Sensing Letters*, 11(1), 19–28. https://doi.org/10.1080/2150704X.2019.1681598

Xue, Z., Yu, X., Liu, B., Tan, X., & Wei, X. (2021). HiResNetAM: hierarchical residual network with attention mechanism for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3566–3580. https://doi.org/10.1109/JSTARS.2021.3069587

Zhong, Z., Li, J., Luo, Z., & Chapman, M. (2017). Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 847–858. https://doi.org/10.1109/TGRS.2017.2755542

Zhong, Y., Hu, X., Luo, C., Wang, X., Zhao, J., & Zhang, L. (2020). WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sensing of Environment*, 250, 112012. https://doi.org/10.1016/j.rse.2020.112012