# Dual-Branch Grouping Multiscale Residual Embedding U-Net and Cross-Attention Fusion Networks for Hyperspectral Image Classification

Ning Ouyang, Chenyu Huang, Leping Lin
School of Information and Communication, Guilin University of Electronic Technology, Guilin, China

***Abstract*—Due to the high cost and time-consuming nature of acquiring labelled samples of hyperspectral data, classification of hyperspectral images with a small number of training samples has been an urgent problem. In recent years, U-Net can train the characteristics of high-precision models with a small amount of data, showing its good performance in small samples. To this end, this paper proposes a dual-branch grouping multiscale residual embedding U-Net and cross-attention fusion networks (DGMRU_CAF) for hyperspectral image classification is proposed. The network contains two branches, spatial GMRU and spectral GMRU, which can reduce the interference between the two types of features, spatial and spectral. In this case, each branch introduces U-Net and designs a grouped multiscale residual block (GMR), which can be used in spatial GMRUs to compensate for the loss of feature information caused by spatial features during down-sampling, and in spectral GMRUs to solve the problem of redundancy in spectral dimensions. Considering the effective fusion of spatial and spectral features between the two branches, the spatial-spectral cross-attention fusion (SSCAF) module is designed to enable the interactive fusion of spatial-spectral features. Experimental results on WHU-Hi-HanChuan and Pavia Center datasets shows the superiority of the method proposed in this paper.**

***Keywords*—U-Net; multiscale; cross-attention; hyperspectral image classification***

## I. INTRODUCTION

Hyperspectral images (HSI) include a wealth of spatial and spectral information [1], which can accurately characterize the physical attributes of features, enhance the ability to discriminate features, and bring great convenience to feature recognition. However, classification of hyperspectral images has its own special problems, such as the redundancy of information in spectral bands [2], the scarcity of training sample data [3], and class imbalance, which bring great challenges to hyperspectral image classification (HSIC).

Traditional HSIC classification methods, such as linear classifier [4], support vector machine [5] and random forest [6], can achieve good classification effect through improvement, but many original traditional methods rely on manual features, and the classification effect is poor when the number of samples is small and the HSI data dimension is high. Therefore, Principal Component Analysis (PCA) has been applied to HSIC by a large number of scholars [7-10]. By compressing the original data to reduce the spectral dimension, the information redundancy between bands and the possible Hughes phenomenon can be avoided, which provides an effective treatment for subsequent feature extraction and enables the network to obtain higher classification accuracy.

With the development of deep learning, the encoder-decoder (U-net) [11] specially designed for biomedical image segmentation has been gradually applied to the field of hyperspectral image classification, which can obtain superior results with less training data. In the absence of datasets, Lin et al. [12] introduced U-Net to solve the problem of complex data capture in practice. Paul et al. [13] combined spectrum partitioning to reduce the redundancy of the spectrum, and then designed U-Net architecture by introducing deep separable convolution to reduce overfitting problems. Besides, due to the clear network structure of U-Net, any customized layer can be easily integrated into the existing network. For example, He et al. [14] embedded the Swin transformer into the classical CNN-based U-Net, which is dedicated to acquiring global contextual information of remote sensing images and obtaining deeper features in the master encoder. Xiao et al. [15] improved the spatial resolution of HSI by fusing spatial features of different scales and depths in the MSI for U-Net.

Moreover, in order to improve the classification performance of hyperspectral images, it has become a major research direction to jointly use spectral and spatial information to design classifiers. The construction of spatial and spectral information through dual branches can make full use of the information. Yang et al. [16] constructed a dual-channel CNN, extracting spectral and spatial information in each channel separately, and then connecting the spatial-spectral features by using cascade, but this simple feature connection cannot capture the complex relationship between the spatial-spectral features. Wang et al. [17] used the grouping strategy and the Long Short-Term Memory (LSTM) model to perceive spectral multi-scale information and obtain spatial context features in spectral finite element and spatial sub-network. Considering the different importance of spectral and spatial components, they used the method of adaptive feature combination for fusion. For effective fusion of spatial-spectral information, Sun et al. [18] designed a weighted self-attention fusion strategy, which combines the output weights of each branch of the previous network with the output weights of self-attention, and obtains efficient fusion on a multi-structured network. Yang et al. [19] used a dual-branch fusion mechanism to promote the exchange of feature information between the two branches through two upstream and downstream modules, so that local fine-grained features could be constructed in more detail and

global context information could be better utilized. These works provide new ideas for dual-branch feature extraction and fusion in HSIC.

In general, the method based on U-Net can better learn the representation of the input natural image, which is conducive to the classification of hyperspectral images to obtain high accuracy and obtain satisfactory results, but some small size information will be lost in the process of down-sampling. the design of fusion mechanism under two-branch conditions will also affect the effectiveness of the network. In this context, we propose a dual-branch grouping multiscale residual embedded U-Net and cross-attention fusion network. Among them, the main contributions are as follows:

- A dual-branching grouping multiscale residual embedded U-Net network (DGMRU) is proposed, which combines grouping multiscale residual block (GMR) and U-Net to extract rich global contextual information and deepen the function of feature network extraction.
- The grouping multiscale residual block (GMR) is constructed for digging multiscale information. The multiscale characteristics of this module enable the network to guide the network to focus on various types of samples at different scales, thereby improving the missed detection problem under spatial features and the redundancy problem under spectral features, and improving the effectiveness of feature extraction.
- A spatial-spectral cross-attention fusion module (SSCAF) is designed to cross-fuse the spatial and spectral features generated by the double branch, that is, to fuse the parameters of the other branch into its own branch, increase the interaction of the two branches, and promote the full fusion of the two branches.

The rest of the paper is organized as follows. Section II describes the general framework of the DGMRU_CAF network, GMR and SSCAF, respectively. Section III discusses the dataset, the experimental settings, the experimental results and the discussion. Finally, in Section IV, conclusions are given.

## II. METHODOLOGY

### A. The Overall Framework of DGMRU_CAF

The DGMRU_CAF proposed in this paper is composed of DGMRU, SSCAF and classification network, as shown in Fig. 1. The DGMRU is divided into a spatial GMRU branch which takes the HSI neighbourhood block $p_n$ as input and a spectral GMRU branch which takes the spectral band $s_n$ as input. Each branch extracts corresponding features from the combined paths of U-Net and GMR with different nuclear scales, so as to obtain deeper feature information. In this regard, the designed GMR enhances the model's perception of multiscale spatial and spectral scales by grouping, multiscaling, and residual connection to retain more detailed feature information. Afterwards, in order to jointly utilize spatial and spectral information, the SSCAF module is constructed. Under the guidance of its own features, the module introduces the features of another branch and carries out interactive fusion to generate spatial-spectral features. Finally, in order to obtain the classification results of HSI, the obtained spatial-spectral features are passed through a classification network consisting of a fully connected layer and a softmax activation layer.

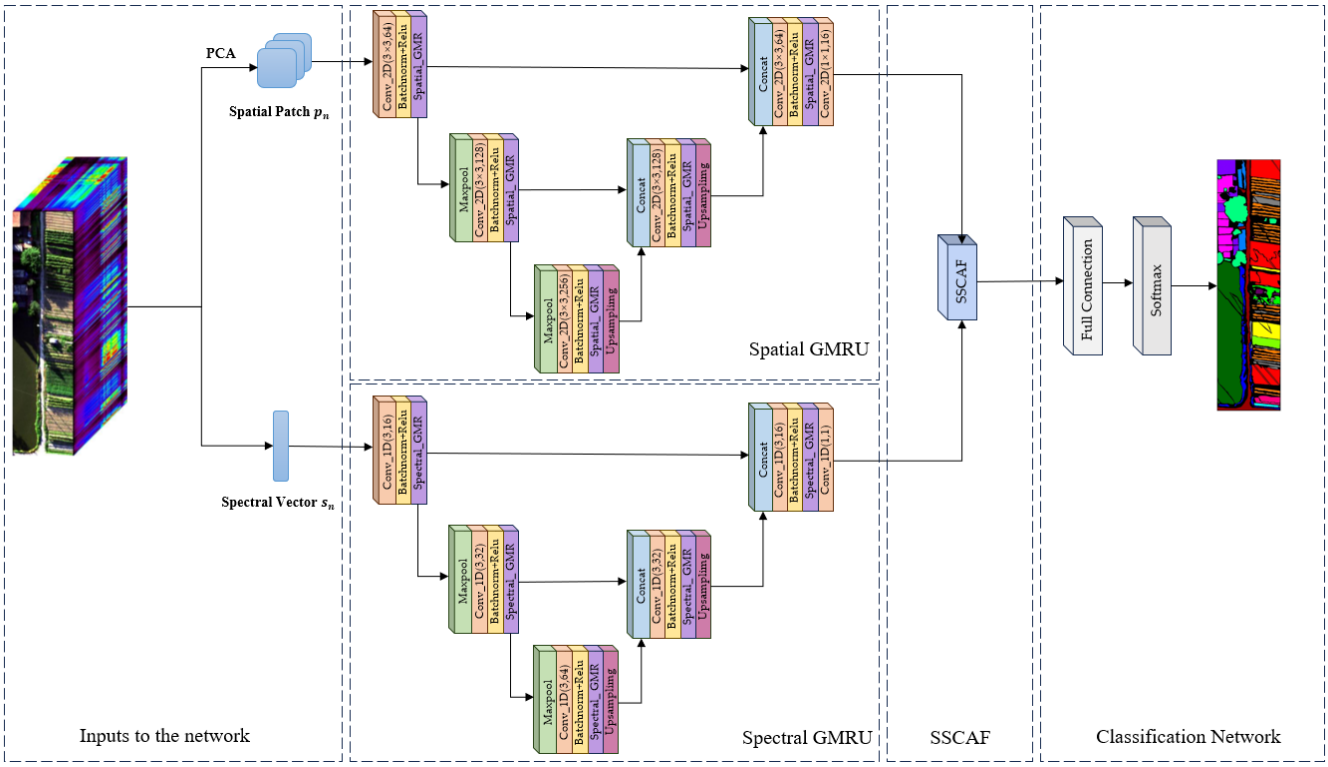



Fig. 1. The overall structure of DGMRU_CAF.

### B. The GMR Module

In this paper, a GMR module is proposed to retain more features without increasing parameters. For each branch, spatial GMR and spectral GMR are designed respectively.

*1)* Spatial GMR: As shown in Fig. 2(a), under the branch of spatial GMRU, for the intermediate features of the input space, spatial GMR uses the grouping module to group its spatial channels in sequence, so that each group of vectors contains different channel information, and each group of spatial vectors is expressed as:

$$c_{spa\_i} = [c_{i\times t}, c_{i\times t+1}, \cdots, c_{i\times t+t}], i = 0, \cdots, \mathrm{g}-1 \quad (1)$$

where, $c_{i\times t}$ is the characteristic information corresponding to the channel in the $i \times t$ segment, t represents the number of channels in each group, and g represents the number of groups.

In the process of down-sampling, the features of small-size objects are easy to be weakened and lost, and it is difficult to recover these features by up-sampling, which leads to the misclassification of small-size objects. In order to solve this problem, this paper uses convolution of different sizes for multi-scale feature extraction after grouping to capture local features inherent in space. The convolution output of each group is:

$$s_i = W_i * x_i + b_i \quad (2)$$

where, $x_i$ is the feature vector of the ith group, $W_i$ is the weight coefficient of the ith group, and $b_i$ is the bias coefficient of the ith group.

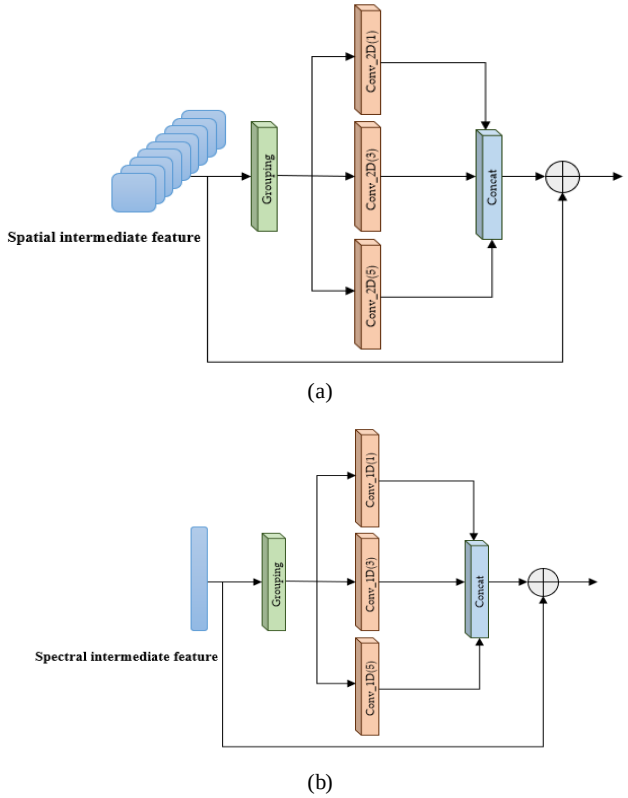



Fig. 2. The structure of GMR module. (a) Spatial GMR (b) Spectral GMR.

Then, in order to complement the context information of features at different scales, all groups are merged by a cascade method. Finally, the rich low-frequency information is transmitted directly through the residual connection, which speeds up the training of the network. In short, using spatial GMR can extract more representative fine features.

*2) Spectral GMR:* Hyperspectral images contain a lot of spectral information, but the spectral information is redundant, which is easy to produce Hughes phenomenon and affect the classification results. In order to cope with this problem, and effectively capture the local relevant information of the spectral band. As shown in Fig. 2(b), for the spectral intermediate features, the grouping module of spectral GMR is used to group their spectral dimensions in sequence, so that each group of spectral vectors contains different spectral band information. Among them, the number of spectrum contained in each group and the distance between spectrum are related to the number of divided groups. Each set of spectral vectors is represented as:

$$r_{spe\_i} = [r_{i\times m}, r_{i\times m+1}, \cdots, r_{i\times m+m}], i = 0, \cdots, \mathrm{g}-1 \quad (3)$$

Where, $r_{i\times m}$ is the characteristic information of the spectral dimension in the $i \times m$ segment, $m$ represents the number of spectral bands in each group, and g represents the number of groups.

After that, convolution of different scales is used to extract the grouped spectral features, so as to weaken the correlation between spectrums and reduce the redundancy of information. After that, the cascade method is used to merge the output spectral features of each group, which complements the local information of the spectral features of different scales and makes full use of the correlation between the spectral bands. Finally, the original global information is propagated directly by residual connection, which alleviates the problem of gradient degradation. In conclusion, the global and local information of spectra can be fully extracted by spectral GMR.

### C. The SSCAF Module

Considering the complementary characteristics between spatial and spectral features, in order to promote the effective fusion of these two types of features, a spatial-spectral cross-attention fusion (SSCAF) module is proposed in this paper. As shown in Fig. 3, the module is a combination of a cross self-attention module, a positional self-attention module (PAM) and a channel self-attention module (CAM). The cross self-attention operation is defined as follows:

$$y_{i,j} = \frac{1}{C(x)} \sum_{\forall i,j} f(f_i, f_j) g(f_i) \quad (4)$$

The $f_i, f_j$ represent the spatial and spectral feature vectors generated by the two branches, respectively, the function $f$ produces the adaptive weight vector between the two vectors, the function $g$ produces the feature representation of the input individual input vectors, and the normalization factor ss is defined as $C(X) = \sum_{\forall i,j} f(f_i, f_j)$.

To further establish internal connections, PAM and CAM modules are introduced to refine spatial and spectral features. Finally, the feature information is summed and

complementarily fused to obtain the final spatial-spectral fusion feature.

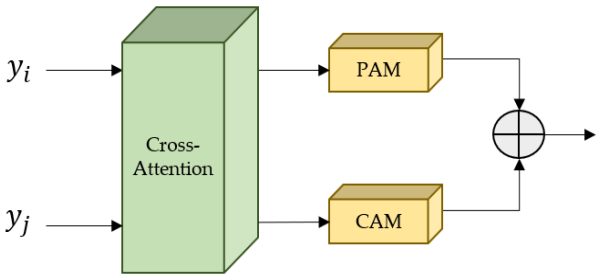



Fig. 3. The structure of SSCAF module.

## III. RESULTS

### A. Dataset and Experimental Setting

In this section, to demonstrate the validity of the proposed method, we conduct a number of experiments on two datasets, which include WHU-Hi-HanChuan (HC) [20],[21] and Pavia Centre (PC). We divided the label samples in different ways for each data set. Table I and Table II provides the specific number of training, test sets and total samples for each class of the two data sets. The false color maps of the two datasets are shown in Fig. 4.

(b)

Fig. 4. False color maps for the two datasets. (a) False color map of HC, (b) False color map of PC.

In the process of training the model, some parameters are set, where the training epoch is set to 200, the batch size is 16, the learning rate is 0.001, the weight decay is 1e-5, and the training is repeated 10 times for all the datasets. In order to prove the superiority of the proposed method, this paper conducts comparative experiments with six advanced methods, namely 2DCNN[22], SSRN[23], A2S2K[24], ASSMN[17], U-Net[11], HyperUnet[13]. The overall accuracy (OA), average accuracy (AA), Kappa coefficient and classification accuracy of single-class are used as the performance evaluation criteria of the model. The higher the each index, the better the classification effect will be.

TABLE I. SAMPLE INFORMATION FOR EACH CLASS IN THE HC DATASET

| Class | Color | Class Name | Train | Test | Total |
|---|---|---|---|---|---|
| 1 | | Strawberry | 200 | 44535 | 44735 |
| 2 | | Cowpea | 200 | 22553 | 22753 |
| 3 | | Soybean | 200 | 10087 | 10287 |
| 4 | | Sorghum | 200 | 5153 | 5353 |
| 5 | | Water Spinach | 200 | 1000 | 1200 |
| 6 | | Watermelon | 200 | 4333 | 4533 |
| 7 | | Greens | 200 | 5703 | 5903 |
| 8 | | Trees | 200 | 17778 | 17978 |
| 9 | | Grass | 200 | 9269 | 9469 |
| 10 | | Red Roof | 200 | 10316 | 10516 |
| 11 | | Gray Roof | 200 | 16711 | 16911 |
| 12 | | Plastic | 200 | 3479 | 3679 |
| 13 | | Bare Soil | 200 | 8916 | 9116 |
| 14 | | Road | 200 | 18360 | 18560 |
| 15 | | Bright Object | 200 | 936 | 1136 |
| 16 | | Water | 200 | 75201 | 75401 |
| Total | | | 3200 | 254330 | 257530 |

TABLE II. SAMPLE INFORMATION FOR EACH CLASS IN THE PC DATASET

| Class | Color | Class Name | Train | Test | Total |
|---|---|---|---|---|---|
| 1 | | Water | 82 | 742 | 824 |
| 2 | | Trees | 82 | 738 | 820 |
| 3 | | Asphalt | 81 | 735 | 816 |
| 4 | | Self-Blocking Bricks | 80 | 728 | 808 |
| 5 | | Bitumen | 80 | 728 | 808 |
| 6 | | Tiles | 100 | 1160 | 1260 |
| 7 | | Shadows | 47 | 429 | 476 |
| 8 | | Meadows | 82 | 742 | 824 |
| 9 | | Bare Soil | 82 | 738 | 820 |
| Total | | | 716 | 6740 | 7456 |

### B. Analysis of Classification Results of Dataset

- Classification Maps and Result of HC Dataset.

The results on the HC dataset are shown in Table III, with the best OA, AA, and Kappa results highlighted in bold. Fig. 5 shows the classification diagram for the different methods.

It can be seen from Table III that our method achieves the best performance, with OA of 96.22%, AA of 96.62%, and Kappa of 95.57%. Compared with other methods, OA, AA, and Kappa are increased by at least 0.8%, 0.76%, and 0.92%. This is because the proposed method has the new idea of combining dual-branch and U-Net, which improves the ability of convolutional feature extraction, so that the method in this paper can achieve the best performance. The grouping multiscale residual block is designed to extract features with different kernel sizes in each group, and reduce the loss of feature information to construct effective feature extraction. The classification results of HSI prove the validity of the method. In addition, it can be seen that the OA of 2DCNN is

the lowest, only 78.91%, which is because 2DCNN is trained only on the spatial dimension, ignoring the information between spectrum, and the model performance is poor. Compared with 2DCNN, U-Net constructs U-shaped network structure, improves classification accuracy and performance, and improves 13.38%, 15.84% and 15.3% respectively in the three evaluation criteria. HyperUnet networks, which combine U-Net and grouping ideas, perform poorly on this dataset, possibly because of poor adaptability to large datasets. In addition, it can be observed that the evaluation value of SSRN is comparable to that of U-Net. SSRN extracts spatial-spectral features through the combination of two continuous spectral blocks and spatial blocks. However, the input of the spatial block comes from the spectral block, which leads to the loss of some spatial information in the spectral block, resulting in poor classification accuracy. The OA of A2S2K is better than that of SSRN, increased by 1.69%, which indicates that the introduction of attention mechanism and adaptive methods has a significant impact on the network. Compared with other single-branch algorithms, the dual-branch ASSMN results in better OA values, which indicates that the full use of spatial and spectral feature information can achieve superior classification results, and the effect is much better than that of single spatial or spectral information. Although the effectiveness of the method in this paper is inferior to other algorithms in some categories, the results of these methods are very close to the results of the best classification, so the OA, AA and kappa coefficients of the method in this paper are the highest among these methods.

From the classification diagram shown in Fig. 5, the "salt and pepper" noise is the most severe because spectral information is not included in 2DCNN, while the classification diagram of other networks shows stronger classification ability because spectral information is taken into account. The method proposed in this paper considers the spatial features of different scales and solves the redundancy problem to obtain more small-size objects and feature information. Therefore, for classification maps with more small sizes, the method proposed in this paper is easier to obtain more accurate and cleaner classification maps, and the classification results of various categories correspond to the results in Table III.

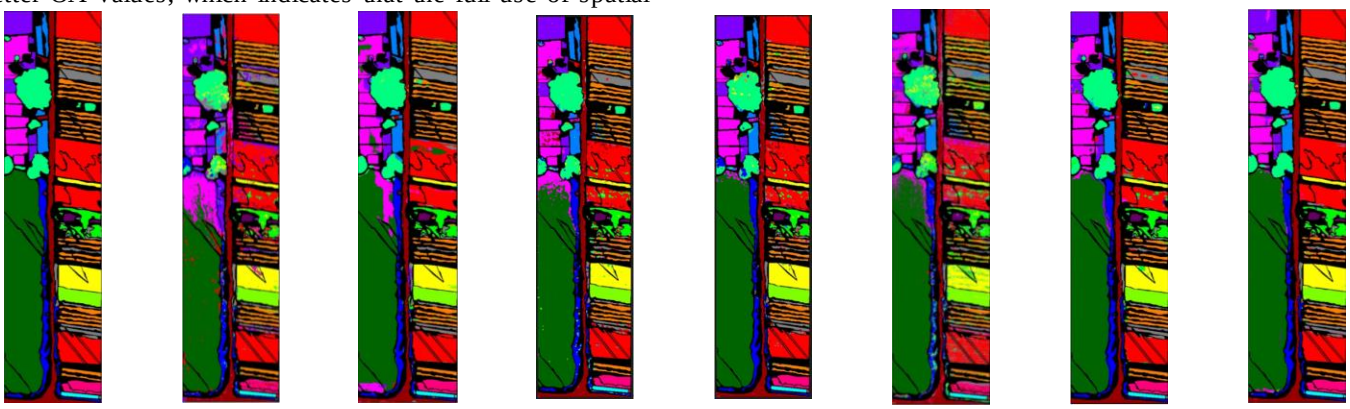



Fig. 5. Classification maps of different methods in HC dataset, (a) Ground truth, (b) 2DCNN, (c) SSRN, (d) A2S2K, (e) U-Net, (f) HyperUnet, (g) ASSMN, (h) Proposed.

TABLE III. CLASSIFICATION RESULTS OF THE HC DATASET

| Class | 2DCNN | SSRN | A2S2K | U-Net | HyperUnet | ASSMN | Proposed |
|---|---|---|---|---|---|---|---|
| 1 | 90.78 | 92.71 | 92.54 | 93.27 | 73.72 | **95.80** | 95.12 |
| 2 | 84.81 | 91.65 | 86.2 | 94.52 | 73.59 | **95.80** | 93.38 |
| 3 | 91.43 | 95.50 | **97.31** | 94.65 | 74.38 | 95.07 | 97.04 |
| 4 | 97.49 | **99.20** | 99.45 | 99.45 | 92.85 | 99.10 | 98.91 |
| 5 | 98.70 | **100** | **100** | **100** | 92.70 | **100** | **100** |
| 6 | 70.69 | 95.5 | 90.53 | 95.19 | 66.74 | **97.71** | 95.47 |
| 7 | 49.04 | 96.35 | 96.82 | **99.35** | 84.69 | 95.75 | 99.10 |
| 8 | 63.24 | 89.83 | 82.07 | 9.17 | 61.84 | 88.77 | **96.38** |
| 9 | 63.02 | 92.08 | **95.46** | 91.74 | 64.89 | 97.85 | 94.47 |
| 10 | 99.46 | 97.37 | **98.92** | 90.25 | 90.92 | 92.54 | 95.86 |
| 11 | 47.87 | 90.88 | **99.15** | 86.80 | 83.32 | 94.57 | 89.52 |
| 12 | 75.88 | 98.93 | 97.64 | 98.47 | 65.24 | **100** | 99.97 |
| 13 | 59.52 | 78.88 | 91.07 | 87.93 | 70.59 | 87.78 | **95.33** |
| 14 | 82.55 | 94.03 | 91.67 | 93.66 | 76.28 | **97.18** | 96.96 |
| 15 | 97.54 | **100** | 99.57 | 97.75 | 84.93 | 99.14 | 99.78 |
| 16 | 80.70 | 92.81 | **98.73** | 90.85 | 94.85 | 96.74 | 98.61 |
| OA（%） | 78.91 | 92.61 | 94.30 | 92.29 | 80.75 | 95.42 | **96.22** |
| AA（%） | 78.29 | 94.11 | 94.83 | 94.13 | 78.22 | 95.86 | **96.62** |
| Kappa×100 | 75.70 | 91.40 | 93.33 | 91.00 | 77.69 | 94.65 | **95.57** |

- Classification Maps and Result of PC Dataset

The results on the PC dataset are shown in Table IV, with the best OA, AA, and Kappa results highlighted in bold. Fig. 6 shows the classification diagram for the different methods.

As shown in Table IV, on this PC dataset, all methods, including 2DCNN, achieve decent classification results. Obviously, the AA of both 2DCNN and SSRN are lower than 90%, which is due to poor classification accuracy in some categories, and the classification accuracy of some categories is less than 80%. The values of A2S3K, U-Net and HyperUnet in OA, AA and Kappa all reach more than 90%, but it is still difficult to improve the classification accuracy for some categories. As a two-branch multi-scale network, ASSMN has more stable classification results. Method of this paper is superior, has the best OA, AA and Kappa evaluation values, and achieves the best accuracy for some specific categories, such as Class 4 Self-Blocking Bricks and Class 5 Bitumen, which further proves its validity in terrain classification.

As shown in Fig. 6, our method is smoother and more consistent.

### C. Ablation Analysis

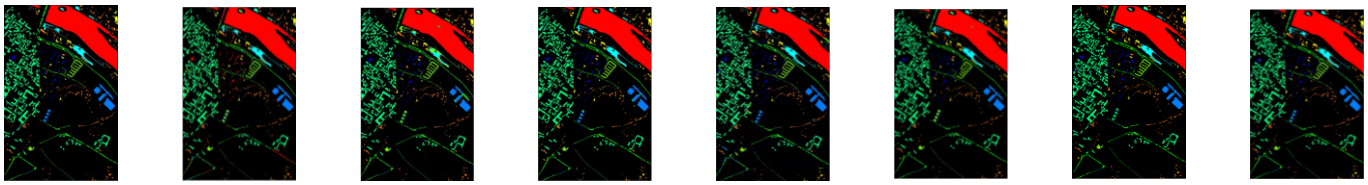
In this part, extensive ablation experiments are conducted to demonstrate the validity of the proposed GMR, SSCAF on the two datasets.

The validity analysis of GMR is shown in Table V. It can be seen that without GMR, the values of OA, AA and Kappa of the model are the lowest in the experiment, because it will lead to some small-size samples being ignored in the process of down-sampling. In contrast, the simultaneous presence of GMR modules with two branches can extract spectral and spatial features more effectively, which contributes to the final classification, and its OA, AA, and Kappa can achieve the best results compared with other comparison strategies. Among them, OA increased by 2.6% and 1.51% in the two datasets, respectively, which means the necessity of GMR. In addition, the OA value of "Only Spe-GMR" is higher than that of "Only Spa-GMR", because the HSI contains enough spectral information to extract more useful feature information from it.

The results of SSCAF ablation experiments are shown in Table VI. It can be found that the integration of SSCAF into the two branches of "With GMR" has significantly improved network performance, which means that SSCAF can complement each other with spatial and spectral information to contribute to the final classification decision. Compared with without SSCAF, OA is increased by 4.54% and 3.03%, AA is increased by 3.14% and 3.59%, and Kappa is increased by 5.32 and 3.46, respectively, which fully proves the necessity of the existence of SSCAF.

TABLE IV. CLASSIFICATION RESULTS OF THE PC DATASET

| Class | 2DCNN | SSRN | A2S2K | U-Net | HyperUnet | ASSMN | Proposed |
|---|---|---|---|---|---|---|---|
| 1 | **100** | 99.77 | **100** | 98.78 | 99.88 | 99.99 | 99.82 |
| 2 | 91.41 | 97.54 | 64.51 | 74.29 | 91.77 | **95.95** | 90.42 |
| 3 | 96.48 | 90.63 | **99.89** | 86.12 | 93.07 | 92.27 | 90.3 |
| 4 | 95.04 | 99.57 | 97.71 | 99.76 | 96.47 | **100** | **100** |
| 5 | 88.54 | 40.29 | 97.56 | 96.94 | 94.30 | 92.82 | **99.09** |
| 6 | 67.07 | 99.14 | **99.95** | 82.65 | 85.05 | 96.55 | 95.74 |
| 7 | 90.05 | 77.97 | 95.71 | **98.35** | 88.24 | 91.45 | 97.16 |
| 8 | 98.75 | 98.14 | 98.34 | 95.68 | **99.65** | 96.50 | 97.78 |
| 9 | 7.49 | **100** | **100** | 87.22 | 99.20 | 95.51 | 97.86 |
| OA（%） | 94.27 | 95.28 | 97.35 | 95.07 | 97.45 | 97.64 | **98.11** |
| AA（%） | 81.65 | 89.23 | 94.85 | 91.09 | 94.18 | 96.01 | **96.48** |
| Kappa×100 | 91.73 | 93.32 | 96.24 | 93.04 | 96.38 | 96.66 | **97.32** |



(**a**) (**b**) (**c**) (d) (**e**) (**f**) (g) (h)

Fig. 6. Classification maps of different methods in PC dataset,(a) Ground truth, (b) 2DCNN, (c) SSRN, (d)A2S2K, (e) U-Net, (f) HyperUnet, (g) ASSMN, (h) Proposed.

TABLE V. EFFECTIVENESS ANALYSIS OF GMR

| Strategy | HC | | | PC | | |
|---|---|---|---|---|---|---|
| | OA | AA | Kappa | OA | AA | Kappa |
| Without GMR | 93.62 | 95.39 | 92.58 | 96.60 | 92.03 | 93.91 |
| Only Spe-GMR | 95.39 | 95.71 | 94.33 | 96.63 | 94.33 | 95.18 |
| Only Spa-GMR | 94.44 | 95.43 | 93.64 | 96.36 | 93.63 | 94.86 |
| With GMR | **96.22** | **96.62** | **95.57** | **98.11** | **96.48** | **97.32** |

TABLE VI. EFFECTIVENESS ANALYSIS OF SSCAF

| Strategy | HC | | | PC | | |
|---|---|---|---|---|---|---|
| | OA | AA | Kappa | OA | AA | Kappa |
| Without SSCAF | 91.68 | 93.48 | 90.25 | 95.08 | 92.89 | 93.86 |
| With SSCAF | **96.22** | **96.62** | **95.57** | **98.11** | **96.48** | **97.32** |

### D. Discussion of Training Times and Testing Times

In order to measure the efficiency of the proposed method, this paper conducted comparative experiments in training and testing time, and the results are shown in Table VII. 2DCNN has the least training time and testing time than other methods, because the simple 2DCNN architecture has fewer training parameters, but the classification accuracy is relatively low. Both U-net and HyperUnet have encoding and decoding path modules, and the addition of more convolutional layers makes the consumption time slightly longer than that of 2DCNN. SSRN and A2S2K use 3D convolution and introduce ResNet, which speeds up convergence and reduces training time. In ASSMN, the combination of dual-branch and multi-scale, together with its strategy of spectrum grouping and spatial grouping, makes the model more complex and requires longer training and testing time. However, the training time and testing time of the method in this paper are average among these comparison methods. The attention mechanism used in the SSCAF module increases the complexity of the proposed network, and the obtained training time and testing time are not the shortest. However, the method proposed in this paper can strike a good balance between accuracy and efficiency, and has certain advantages.

TABLE VII. RUNNING TIME OF DIFFERENT METHODS ON TWO DATASETS

| Dataset | | 2DCNN | SSRN | A2S2K | U-Net | HyperUnet | ASSMN | Proposed |
|---|---|---|---|---|---|---|---|---|
| HC | Train(s) | 8.96 | 48.53 | 63.96 | 22.54 | 38.05 | 192.15 | 82.46 |
| | Test(s) | 42.15 | 123.80 | 205.50 | 61.27 | 82.04 | 121.01 | 138.05 |
| PC | Train(s) | 3.11 | 7.84 | 11.75 | 7.45 | 8.65 | 54.90 | 18.88 |
| | Test(s) | 9.32 | 18.74 | 23.53 | 34.74 | 23.30 | 69.78 | 73.41 |

## IV. CONCLUSION

In this paper, we propose a dual-branch grouping multiscale residual embedded U-Net and cross-attention fusion network for hyperspectral image classification to improve the classification accuracy in the presence of sparse training samples. The designed DGMRU module is used to extract multiscale context information feature, which is suitable for the case of insufficient HSI samples. Among them, the designed GMR module increases the receptive field without adding parameters, and the feature extraction effect is better than that of the non-existent GMR module, which proves the necessity of this module. In addition, the proposed SSCAF maximizes the utilization of spatial-spectral features by constructing the intrinsic relationship between spatial and spectral features through cross-attention. Compared with other advanced algorithms, the method proposed in this paper has the best experimental results, and in the two data sets, OA increases by 0.8% and 0.47% at least, which is feasible and effective. In the future, we will consider further reducing the complexity of the network model and improving the computational efficiency while maintaining the classification accuracy.

## ACKNOWLEDGMENT

The authors thank other laboratories for providing hyperspectral datasets, as well as editors and anonymous reviewers for their comments and suggestions. This work was supported by the National Natural Science Foundation of China (No.62001133).

## REFERENCES

[1] X. Li, Z. Li, H. Qiu, G. Hou, and P. Fan, "An overview of hyperspectral image feature extraction, classification methods and the methods based on small samples," Appl. Spectrosc. Rev., vol. 58, no. 6, pp. 367–400, Jul. 2023.

[2] R. N. Patro, S. Subudhi, P. K. Biswal, and F. Dell'acqua, "A Review of Unsupervised Band Selection Techniques: Land Cover Classification for Hyperspectral Earth Observation Data," IEEE Geosci. Remote Sens. Mag., vol. 9, no. 3, pp. 72–111, Sep. 2021.

[3] X. Wang, J. Liu, W. Chi, W. Wang, and Y. Ni, "Advances in Hyperspectral Image Classification Methods with Small Samples: A Review," Remote Sens., vol. 15, no. 15, Art. no. 15, Jan. 2023.

[4] M. Shambulinga and G. Sadashivappa, "Supervised hyperspectral image classification using svm and linear discriminant analysis," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 10, pp. 403–409, 2020.

[5] J. Dr. J. H. Harikiran, "Hyperspectral image classification using support vector machines," IAES Int. J. Artif. Intell. IJ-AI, vol. 9, no. 4, p. 684, Dec. 2020.

[6] F. Tong and Y. Zhang, "Exploiting Spectral–Spatial Information Using Deep Random Forest for Hyperspectral Imagery Classification," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[7] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," IETE Tech. Rev., vol. 38, no. 4, pp. 377–396, 2021.

[8] H. Fu, G. Sun, J. Ren, A. Zhang, and X. Jia, "Fusion of PCA and Segmented-PCA Domain Multiscale 2-D-SSA for Effective Spectral-Spatial Feature Extraction and Data Classification in Hyperspectral Imagery," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022.

[9] Q. Liu, D. Xue, Y. Tang, Y. Zhao, J. Ren, and H. Sun, "PSSA: PCA-Domain Superpixelwise Singular Spectral Analysis for Unsupervised Hyperspectral Image Classification," Remote Sens., vol. 15, no. 4, p. 890, 2023.

[10] X. Zhang, X. Jiang, J. Jiang, Y. Zhang, X. Liu, and Z. Cai, "Spectral–Spatial and Superpixelwise PCA for Unsupervised Feature Extraction of Hyperspectral Imagery," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–10, 2022.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[12] M. Lin, W. Jing, D. Di, G. Chen, and H. Song, "Context-aware attentional graph U-Net for hyperspectral image classification," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2021.

[13] A. Paul and S. Bhoumik, "Classification of hyperspectral imagery using spectrally partitioned HyperUnet," Neural Comput. Appl., pp. 1–10, 2022.

[14] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–15, 2022.

[15] J. Xiao, J. Li, Q. Yuan, and L. Zhang, "A dual-UNet with multistage details injection for hyperspectral image fusion," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2021.

[16] J. Yang, Y. Zhao, J. C.-W. Chan, and C. Yi, "Hyperspectral image classification using two-channel deep convolutional neural network," in 2016 IEEE international geoscience and remote sensing symposium (IGARSS), IEEE, 2016, pp. 5079–5082.

[17] D. Wang, B. Du, L. Zhang, and Y. Xu, "Adaptive spectral–spatial multiscale contextual feature extraction for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 3, pp. 2461–2477, 2020.

[18] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multi-structure KELM with attention fusion strategy for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–17, 2022.

[19] L. Yang et al., "FusionNet: a convolution–transformer fusion network for hyperspectral image classification," Remote Sens., vol. 14, no. 16, p. 4066, 2022.

[20] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "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 Sens. Environ., vol. 250, p. 112012, 2020.

[21] Y. Zhong et al., "Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications," IEEE Geosci. Remote Sens. Mag., vol. 6, no. 4, pp. 46–62, 2018.

[22] X. Yang, Y. Ye, X. Li, R. Y. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9, pp. 5408–5423, 2018.

[23] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 847–858, 2017.

[24] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 9, pp. 7831–7843, 2020.