Hyperspectral Image Classification With Multiattention Fusion Network

Zhaokui Li, Xiaodan Zhao, Yimin Xu, Wei Li, *Senior Member, IEEE*, Lin Zhai,
Zhuoqun Fang, and Xiangbin Shi

Abstract—**Hyperspectral image (HSI) has hundreds of continuous bands that contain a lot of redundant information. Besides, a spatial patch of a hyperspectral cube often contains some pixels different from the center pixel category, which are usually called interference pixels. The existence of such interference pixels has a negative effect on extracting more discriminative information. Therefore, in this letter, a multiattention fusion network (MAFN) for HSI classification is proposed. Compared with the current state-of-the-art methods, MAFN uses band attention module (BAM) and spatial attention module (SAM), respectively, to alleviate the influence of redundant bands and interfering pixels. In this way, MAFN realizes feature reuse and obtains complementary information from different levels by combining multiattention and multilevel fusion mechanisms, which can extract more representative features. Experiments were conducted on two public HSI data sets to demonstrate the effectiveness of MAFN. Our source code is available at https://github.com/Li-ZK/MAFN-2021.**

Index Terms—**Band attention, hyperspectral image (HIS) classification, multilayer feature fusion, spatial attention.**

I. INTRODUCTION

HYPERSPECTRAL image (HSI) contains hundreds of narrow spectral bands with extremely high spectral resolution. Therefore, HSI classification is widely used in various remote sensing applications, such as urban development, environmental monitoring, military reconnaissance, and land cover classification.

In the past two decades, a large number of methods have been proposed to solve the HSI classification task. Those early methods mainly adopted machine learning techniques, such as k-nearest neighbor (KNN) [1], support vector machine [2], sparse representation [2], and multinomial logistic regression [3]. However, these classifiers often perform poorly since they are not applicable in highly nonlinear hyperspectral remote sensing data. To address this problem, deep learning

models were introduced to HSI classification, in recent years, and achieved better performance.

In the early stage of deep learning, some typical deep feature extractors based on spectral features were proposed, such as stacked autoencoder (SAE) [4], recurrent neural network (RNN) [4], and deep belief network (DBN) [5]. However, due to the influence of the redundant bands of HSI and the neglect of spatial information, the classification results are not ideal. To make better use of both spectral and spatial information, Yang *et al.* [6] proposed a two-brance convolutional neural network (TWO-CNN). However, this method used 1-D convolution (1-D CONV) to directly extract spectral information, resulting in the loss of spatial information, which is not conducive to the improvement of classification accuracy. In addition, a supervised spectral–spatial feature extractor based on the residual network was proposed in the literature [7]. However, this method put spatial learning after spectral learning, destroyed the original characteristics of spatial information, and ignored the correlation between spatial information and spectral information. Furthermore, feature fusion is helpful in the improvement of HSI classification. Song *et al.* [8] designed a deep feature fusion network that can extract features at different levels. However, the feature extraction after fusion had been neglected.

More recently, the attention mechanism has been used extensively in HSI classification. To solve the problem of band redundancy, Dong *et al.* [9] proposed a band attention module (BAM) and applied it in visual geometry group network (VGGNet). However, this letter used high-level abstract features and ignored the important features of the shallow layer. Moreover, a spatial attention mechanism [10] that can suppress the influence of interfering pixels was studied. However, just like spectral-spatial residual network (SSRN), the spatial feature extraction sequence was placed behind the spectral feature extraction, which destroyed the correlation between spectral and spatial information. In addition, Li *et al.* [11] proposed a network double-branch dual-attention (DBDA) combining both spectral attention mechanism and spatial attention mechanism. However, the attention mechanism used in this method is located at the back end of the network, so it does not play its role in the beginning.

To further improve the classification accuracy and suppress the interference pixels in a 3-D cube, multiattention fusion network (MAFN) is proposed. MAFN adopts BAM to address the problem of band redundancy and improve the overall performance. SAM is used to extract the regional features consistent with the central pixel category, thereby extracting more distinguishing features. In addition, multilevel feature fusion
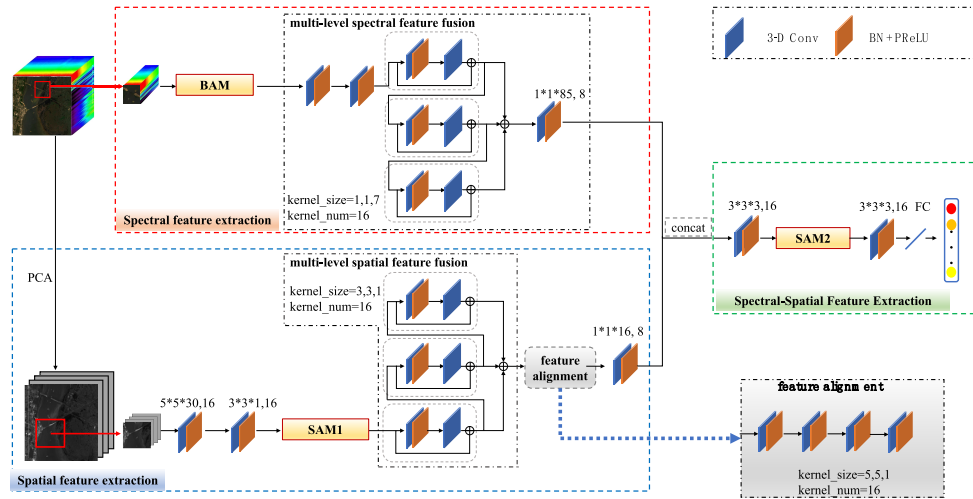
Fig. 1.   Overall flowchart of MAFN.

is also used in this letter to extract different features of low, middle, and high levels. Finally, by combining multiattention and multilevel feature fusion mechanisms, MAFN achieves feature reuse and obtains supplementary information from different levels so as to extract more representative features.

To summarize, our main contributions of this letter can be condensed into two parts. First, the spectral and spatial attention mechanisms are used, respectively, to alleviate the influence of redundant bands and interfering pixels. Second, more representative multiattention fusion features are extracted through the combination of multiattention and multilevel fusion mechanism, which can realize feature reuse and obtain complementary information from different levels.

II. Proposed MAFN Framework

In this section, we describe the proposed MAFN method. The overall flowchart of MAFN is shown in Fig. 1. As shown in Fig. 1, the MAFN consists of three main components: spectral feature extraction, spatial feature extraction, and joint spectral–spatial feature extraction. To better illustrate the overall flow of MAFN, a widely used data set Kennedy Space Center (KSC) is used as an example input in this section. In the rest of this section, these three components are introduced in detail.

A. Spectral Feature Extraction

The spectral feature extraction part (red dashed box) is used to extract more representative spectral features. First, multiple 3-D cubes with shape $7 \times 7 \times 176$ are sampled from the HSI and regarded as an input of this channel. Those input cubes are first passed to BAM to weigh hundreds of bands. After two layers of shallow convolution, three consecutive residual structures are used to learn more distinguishing features. Those features extracted from the three residual blocks are fused at low, medium, and high levels. Combining the attention module with multilevel fusion, the attention module can be reused to extract spectral attention information.

BAM is used to mitigate band redundancy in HSI. The structure of BAM is shown in Fig. 2. First, a series of 2-D convolutions are used to expand the receptive field, while pooling is used to reduce the resolution. In this way, global
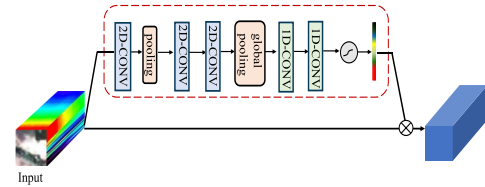

Fig. 2.   Structure of BAM.

TABLE I
PARAMETER TABLE OF BAM

2-D Conv	16, (3, 3)
Pooling	(2,2)
2-D Conv	32, (3, 3)
2-D Conv	32, (3, 3)
1-D Conv	channel/r, (32)
1-D Conv	channel, (channel/r)

information of the space domain is obtained. Next, global average pooling sums up the spatial information. After that, two 1-D convolutions are used to learn the nonlinear relationship among the bands. A parameter (i.e., r) is used to control the parameters of the 1-D convolution kernel, thereby controlling the degree of aggregation of spectral information. Finally, after using the sigmoid activation function, a weighted vector (200 elements in length) with global spectral information is obtained. By performing an inner product operation on the weighted vector and the input 3-D cube, more representative spectral features are obtained. Using BAM can weight spectral information, thereby reducing the impact of redundant bands on the classification results. The parameter table of BAM is shown in Table I.

B. Spatial Feature Extraction

The input patch size of spatial feature extraction (blue dashed box) is larger, which contains more spatial information. To reduce the computational cost, the principal component analysis (PCA) algorithm is first performed on HSI. First, multiple cubes shaped like $27 \times 27 \times 30$ are used as an input for spatial feature extraction. Then, two layers of convolution are used to reduce the number of channels and the amount
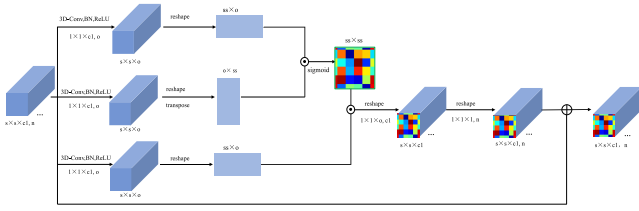
Fig. 3.   Structure of SAM.

Fig. 4.   OA on the two data sets at different r in BAM.

of calculation. Afterward, the spatial features are extracted through SAM and three residual blocks. Next, a feature alignment operation is performed, which can reduce the dimension of the spatial feature to the same as the spectral branch to wait for the fusion with the spectral feature.

In each patch, there are pixels with different labels from the center pixel, which will interfere with the classification results. Therefore, SAM is used in MAFN, whose purpose is to suppress the influence of interfering pixels on the classification result, so that the network pays more attention to the pixels with the same label as the central pixel when extracting features.

The structure of SAM is shown in Fig. 3. $X_k \in \mathbb{R}^{s \times s \times c_1}$ represents the input of the module, which means that the input spatial scale is $s \times s$ and the number of bands is c_1. The 3-D convolution with $1 \times 1 \times c_1$ kernels is used to convert the input into three feature maps of $s \times s \times c_1$. These three feature maps are represented by $f(X_k) \in \mathbb{R}^{s \times s \times o}$, $g(X_k) \in \mathbb{R}^{s \times s \times o}$, and $h(X_k) \in \mathbb{R}^{s \times s \times o}$ from top to bottom. The purpose of this step is to reduce the number of channels and simplify calculations. Among them, $f(X_k^l)$ is calculated as follows F:

$$f(X_k) = \sigma\left(W_f * X_k + b_f\right) \tag{1}$$

where W_f and b_f, respectively, represent the weight parameter and the bias parameter in the convolution operation. Similarly, $g(X_k^l)$ and $h(X_k^l)$ are calculated as follows:

$$g(X_k) = \sigma\left(W_g * X_k + b_g\right) \tag{2}$$

$$h(X_k) = \sigma\left(W_h * X_k + b_h\right). \tag{3}$$

Next, three feature maps obtained in the previous step are reshaped to $ss \times o$, and $f(X_k)$ is multiplied by $g(X_k)^T$

$$R = f(X_k)g(X_k)^T. \tag{4}$$

This step is to calculate the correlation between any two pixels in the feature map. After performing the normalization operation using the sigmoid activation function, a mask weighted for all pixels can be obtained, and the area with a larger weight is more important. Then, as shown in formula (5), the attention coefficient is correspondingly multiplied by the feature map $h(X_k)$ to generate attention features

$$\text{Att} = Rh(X_k). \tag{5}$$

Finally, the number of channels is expanded by convolution to get the same size as the input. To facilitate the convergence of the algorithm, a skip connection is performed to add the attention feature Att to the input feature X_k.

Like the spectral branch, the spatial branch also uses low-, medium-, and high-level feature fusion, which can reuse the spatial attention features extracted from the low level to
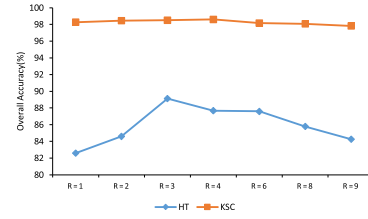
achieve the complementary advantages of multiple features and obtain more robust and accurate results.

C. Joint Spectral–Spatial Feature Extraction

The green dashed box in Fig. 1 is the extraction of joint spectral–spatial features. The features extracted from the spatial and spectral branches are concatenated and then subjected to a $3 \times 3 \times 3$ convolution operation. Since the cascade of the two branches introduces in some interference pixels from the spectral branch, the fused feature is fed into the SAM2 again to extract the features of the spatial attention region. Finally, the softmax classifier is used to obtain the classification results. There is no real difference between SAM1 and SAM2. The numbers 1 and 2 are only used to distinguish between the two uses of SAM.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Data Sets and Experimental Settings

Two public hyperspectral data sets are selected to prove the effectiveness of our method. The Houston 2013 (HT) data set is the competition data of the 2013 GRSS Data Fusion contest, which describes the landscape of Houston University and its surrounding areas. The size of the data set is 349×1905, and the spatial resolution is 2.5 m per pixel. The data set contains 144 spectral bands and 15 kinds of surface features. The KSC data set was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensors in Florida. The data set contains 13 ground-truth classes with a spatial resolution of 1.8 m, containing a total of 512×614 pixels. After removing the noisy bands, there are 176 bands for study.

In the experiments, the training samples and test samples of the HT data set are fixed, which are 2832 and 12,197, respectively. Moreover, using Adam as the optimizer, the learning rate is 0.001. For KSC, 5% of samples are selected as the training set, and stochastic gradient descent (SGD) optimizer is used, and the learning rate is set to 0.0005. The batch size is chosen to be 16. Also, 50% and 60% dropout operations are adopted, respectively. Three evaluation criteria, average accuracy (AA), overall accuracy (OA), and Kappa coefficient (K), are used to evaluate the performance of MAFN.

B. Analysis of Parameters

1) Influence of Parameter r in BAM: As described in BAM, two 1-D-Conv of parameter r control the aggregation of spectral information. In this part, MAFN is evaluated when $r = 1, 2, 3, 4, 6, 8$, and 9. Fig. 4 shows the OA of different data sets with different r values. Obviously, when $r = 4$, MAFN achieves the highest overall accuracy in the KSC data set. When $r = 3$, the classification effect of the HT data set is the best.
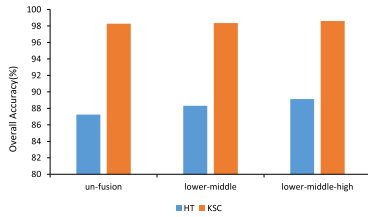
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

4

IEEE GEOSCIENCE AND REMOTE SENSING LETTERS

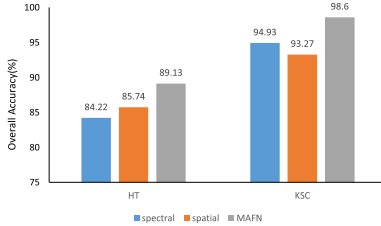Fig. 5. Validation of multi-level fusion.

Fig. 6. Validation of dual-branch network.

2) Validation of Multilevel Feature Fusion: For deep networks, features at different levels have different meanings. Low-level features have higher resolution and more details, but contain lower semantics and more noise. In contrast, high-level features have stronger semantic information, but the resolution is low, and the perception of details is poor. Therefore, in this letter, we did comparative experiments on "un-fusion," "low-middle-level fusion," and "low-middle-high-level fusion." From the column chart in Fig. 5, the network with "low-middle-high-level fusion" outperforms the other two on both the HT and KSC data sets. These results prove the effectiveness of the multilevel fusion, which contributes to higher classification accuracy.

3) Validation of Dual-Branch Network: We carried out comparative experiments of spectral feature extraction (spectral), spatial feature extraction (spatial), and spectral–spatial feature extraction. The experimental results can be showed vividly from the histogram in Fig. 6. The performance of spectral feature extraction and spatial feature extraction is not good. Only by extracting spectral–spatial features simultaneously can better classification results be obtained.

4) Validation of Attention Module: The experiment in this section is to prove the effectiveness of the various attention modules. It can be observed from Table II that we sequentially compared the basic network (basic net), the network using only BAM, the network that uses SAM only in spatial feature extraction (SAM1), the network using SAMs in both spatial feature extraction and joint spectral spatial feature extraction (SAM1 + SAM2), the network using attention modules in both spectral and spatial feature extraction (BAM + SAM1), and the final method (MAFN) on AA, OA, and K. On the HT data set, it can be seen that OA of MAFN is 4.04% higher than the basic net. The effect of other intermediate networks achieves higher accuracy compared with the basic network, which also proves the effectiveness of the attention mechanism. There is also a 2% improvement between MAFN and basic net on the KSC data set. This result proves the effectiveness of the three attention modules in MAFN.

TABLE II
VALIDATION IN MAFN OF EACH ATTENTION MODULE

	Metric Model	OA(%)	AA(%)	K×100
HT	basic net	85.09±0.87	85.07±1.25	83.95±0.96
	BAM	85.87±0.63	87.34±0.67	87.72±0.99
	SAM1	85.69±0.31	87.79±1.01	84.49±0.34
	SAM1+SAM2	87.27±0.87	88.16±0.49	86.41±1.29
	BAM+SAM1	87.12±0.95	87.86±1.21	87.17±0.60
	MAFN	**89.13±1.37**	**90.31±1.84**	**88.36±1.57**
KSC	basic net	96.63±0.25	95.73±0.50	96.31±0.48
	BAM	97.43±0.12	96.67±0.38	97.09±0.19
	SAM1	97.37±0.28	96.48±0.42	97.10±0.34
	SAM1+SAM2	98.30±0.24	97.46±0.44	98.12±0.27
	BAM+SAM1	97.63±0.28	96.89±0.56	97.37±0.32
	MAFN	**98.60±0.62**	**98.02±0.57**	**98.42±0.38**

TABLE III
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS
ON HT DATA SET

	TWO-CNN	BAM-CM	SSAN	DFFN	SSRN	DBDA	MAFN
1	71.50	68.51	82.08	66.52	**100**	88.48	96.75
2	83.52	88.48	81.65	88.34	81.68	85.06	**90.32**
3	10.04	15.53	10.20	66.95	**100**	**100**	**100**
4	73.53	81.50	96.57	95.28	**99.34**	95.47	93.98
5	81.39	79.57	76.19	98.62	**100**	**100**	96.77
6	41.93	67.77	30.26	94.94	**100**	16.15	90.15
7	64.61	71.45	74.01	93.19	95.23	92.21	**98.08**
8	66.89	70.14	55.82	86.55	96.05	77.10	**97.32**
9	60.46	69.54	85.19	87.66	64.11	**96.57**	93.88
10	55.09	71.49	54.53	59.34	**85.68**	83.95	68.93
11	52.05	59.90	71.13	78.77	**98.93**	92.99	90.18
12	39.45	49.19	35.86	64.34	50.79	**95.03**	80.18
13	34.57	49.54	63.32	27.48	49.42	**99.17**	93.82
14	76.40	61.66	75.91	68.05	75.00	97.51	**98.73**
15	78.09	75.35	90.26	71.74	**99.54**	95.78	88.38
OA	56.85 ± 0.47	61.87 ± 0.24	62.54 ± 0.64	77.40 ± 0.74	84.81 ± 0.59	86.53 ± 1.21	**89.13** ±1.37
AA	58.67 ± 0.18	64.98 ± 0.85	67.54 ± 0.71	79.78 ± 1.02	86.53 ± 1.38	88.59 ± 1.59	**90.31** ±1.84
K×100	53.35 ± 0.50	58.80 ± 0.27	61.58 ± 0.53	75.21 ± 0.81	83.69 ± 0.74	85.45 ± 1.31	**88.36** ±1.57

C. Experimental Results

The proposed MAFN is evaluated with TWO-CNN [6], SSRN [7], deep feature fusion network (DFFN) [8], BAM-classification model (BAM-CM) [9], spectral-spatial attention networks (SSAN) [10], and DBDA [11]. The HT data set is different from KSC in that it has a fixed training set and test set, which is challenging for the generalization of the model. The comparison results on the HT data set are shown in Table III. From the perspective of OA, AA, and K, the classification results of MAFN have a relatively stable improvement compared with other methods. From the classification results of each category, except for the 10th category—highway, other classification results have achieved satisfactory results. This proves that MAFN can extract more effective features, and the network has more robustness and generalization ability, which is helpful in improving the classification results. The classification map is shown in Fig. 7. For the KSC data set, the data distribution is relatively uniform.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

LI *et al.*: HSI CLASSIFICATION WITH MAFN

5

TABLE IV
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS
ON KSC DATA SET

|	TWO-CNN	BAM-CM	SSAN	DFFN	SSRN	DBDA	MAFN
1	91.39	93.45	91.41	91.69	98.49	98.91	**99.07**
2	75.25	96	93.01	93.5	**100**	96.27	**100**
3	64.46	72.81	95.47	85.58	95.61	90.24	**97.79**
4	63.24	97.67	95.78	92.61	96.23	95.02	**99.51**
5	74.1	87.20	97.08	93.66	74.11	98.49	**98.51**
6	56.66	87.62	90.27	94.73	**99.51**	96.41	98.99
7	60.58	**100**	93.25	96.51	**100**	94.04	95.5
8	69.13	84.31	96.28	92.58	96.91	95.13	**98.19**
9	78.99	85.89	83.33	89.91	**100**	92.69	93.13
10	86.75	90.96	93.78	**100**	**100**	98.83	**100**
11	93.58	94.28	**100**	97.54	99.74	99.72	**100**
12	97.91	97.65	**100**	**100**	**100**	**100**	**100**
13	92.24	96.13	**100**	99.37	**100**	**100**	**100**
OA	83.97	91.23	94.49	94.26	97.68	97.70	**98.60**
	±0.64	±0.56	±0.36	±0.52	±0.43	±0.94	±0.62
AA	79.95	91.08	95.61	93.65	97.17	95.42	**98.02**
	±1.57	±0.35	±0.31	±0.44	±0.36	±1.66	±0.57
K×100	82.36	90.23	93.76	93.85	97.23	97.08	**98.42**
	±0.80	±0.55	±0.41	±0.37	±0.30	±1.05	±0.38
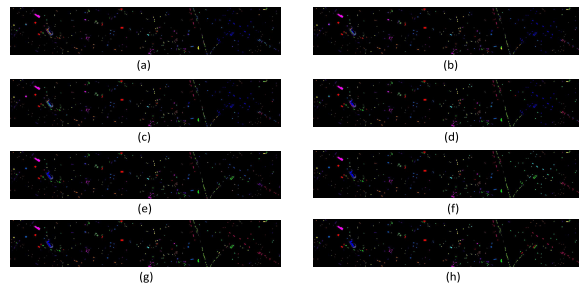
Fig. 7. Classification maps on the HT data set (a) TWO-CNN, (b) BAM-CM, (c) SSAN, (d) DFFN, (e) SSRN, (f) DBDA, (g) MAFN, and (h) ground-truth.
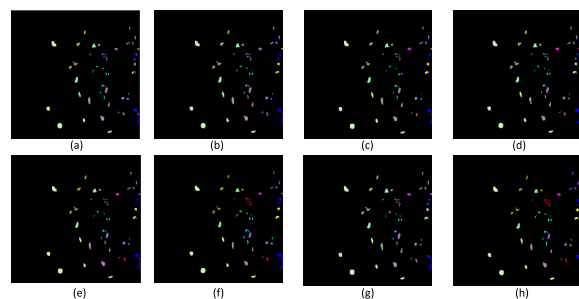

Fig. 8. Classification maps on the KSC data set (a) TWO-CNN, (b) BAM-CM, (c) SSAN, (d) DFFN, (e) SSRN, (f) DBDA, (g) MAFN, and (h) ground-truth.

As shown in Table IV, the classification results of MAFN in categories 6, 7, and 9 are slightly lower than SSRN. But on the whole, the classification result of our method is smoother, and the OA value is nearly 1% higher than SSRN. Compared with the other methods, MAFN steadily wins the results of each class and the final result. The classification map of each method on KSC is shown in Fig. 8. The experimental results on the two data sets can prove the effectiveness of the multiattention mechanism and multilevel fusion.

In addition, we also carried out experiments on samples with different proportions of KSC data sets. In Fig. 9, as the training samples increase from 1% to 5%, the accuracy of each method steadily improves. At the same time, it is obvious that
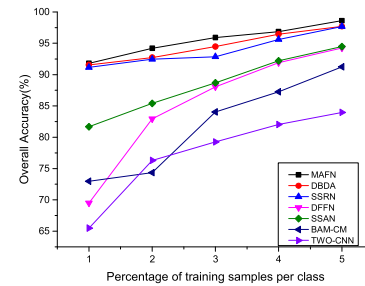

Fig. 9. OA of 1%–5% percentage of training samples by all methods on the KSC data set.

the polyline of MAFN is always above other methods, which demonstrate the effectiveness of the proposed MAFN. This phenomenon shows that MAFN is more adaptive when few training samples are available.

IV. CONCLUSION

In this research, a novel model MAFN was proposed for the HSI classification. BAM was used to alleviate band redundancy problems, while SAM was used to suppress the influence of interfering pixels. The combination of multi-attention and multilevel fusion mechanisms realized feature reuse and obtained complementary information from different feature levels, which can extract more representative features. The comparison experiments with other methods on the HT and KSC data sets demonstrated effectiveness of the proposed MAFN.

REFERENCES

[1] L. Samaniego, A. Bardossy, and K. Schulz, "Supervised classification of remotely sensed imagery using a modified *K*-NN technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2112–2125, Jul. 2008.

[2] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[3] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[4] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[5] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[6] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[7] 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, Feb. 2018.

[8] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[9] H. Dong, L. Zhang, and B. Zou, "Band attention convolutional networks for hyperspectral image classification," 2019, *arXiv:1906.04379*. [Online]. Available: http://arxiv.org/abs/1906.04379

[10] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[11] 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, p. 582, Feb. 2020.