DARecNet-BS: Unsupervised Dual-Attention Reconstruction Network for Hyperspectral Band Selection

Swalpa Kumar Roy, *Student Member, IEEE*, Sayantan Das, Tiecheng Song, *Member, IEEE*, and Bhabatosh Chanda

***Abstract*— Due to the existence of noise and spectral redundancies in hyperspectral images (HSIs), the band selection (BS) is highly required and can be achieved through the attention mechanism. However, existing BS methods fail to consider global interaction between the spectral information and spatial information in a nonlinear fashion. In this letter, we propose an end-to-end unsupervised dual-attention reconstruction network for BS (`DARecNet-BS`). The proposed network employs a dual-attention mechanism, i.e., position attention module (PAM) and channel attention module (CAM), to recalibrate the feature maps and subsequently uses a 3-D reconstruction network to restore the original HSI. This way, the long-range nonlinear contextual information in spectral and spatial directions is captured, and the informative band subset can be selected. Experiments are conducted on three well-known hyperspectral data sets, i.e., Indian Pines (IP), University of Pavia (UP), and Salinas (SA), to compare existing BS approaches, and the proposed `DARecNet-BS` can effectively select less redundant bands with comparable or better classification accuracy. The source code will be made publicly available at https://github.com/ucalyptus/DARecNet-BS.**

***Index Terms*— Band selection (BS), channel attention, hyperspectral images (HSIs), position attention.**

I. Introduction

HYPERSPECTRAL images (HSIs) contain rich information on a wide range of continuous narrow spectral bands with a high spatial resolution and have been extensively studied in image processing and computer vision applications [1]–[3]. Due to a large number of spectrum bands present in the data, HSI always suffers from "the curse of dimensionality" and a huge computational cost. To tackle this problem, it is crucial to select the most informative spectral bands so that the characteristic of the data is well preserved.

Two types of dimensionality reduction techniques, i.e., feature extraction and feature selection, are widely used to analyze HSI. Feature extraction aims to find a mapping from the original high-dimensional features to a low-dimensional space typically using subspace learning [4], [5] or averaging based methods [6], while feature selection aims to represent the original data by selecting the most informative subsets. Compared with feature extraction methods, band selection (BS) approaches [7] can better represent the physical information of the original data and, thus, can be easily adopted in practice.

The BS techniques can be categorized into supervised and unsupervised methods [8]. Due to the lack of proper ground truth and robust performance, unsupervised methods have received a lot of attention in the last few decades. Among them, the searching-, clustering-, and ranking-based methods are commonly used for BS. In searching based methods, a combination of objective functions is arranged and optimized using, e.g., a time-consuming heuristic search [9]. In clustering-based approaches, the similarities among different bands are found by performing suitable clustering algorithms, such as subspace clustering [10] and sparse nonnegative matrix factorization (SNMF) [11]. The ranking-based approaches find informative spectral bands by assigning weight or rank for each spectral band based on the estimated significance, such as sparse representation (SpaBS) [12] and geometry-based BS (OPBS) [13].

Recently, deep neural networks have received much attention in vision research due to their hierarchical representation ability and good generalization ability, which have been successfully adopted in the HSI domain [1], [2], [14]–[17]. This inspired researchers to develop various attention mechanisms that not only suggest where to focus but also improve the feature representation quality [18], [19]. The attention block finds meaningful patterns by dynamically extracting feature maps to help the classification and, meanwhile, suppressing ineffective feature maps to reduce the misclassification probability. The band or channel attention module (CAM) was initially introduced in BS network (BSNet-Conv) [20] to select the majority of spectral bands carrying useful information for classification. However, BSNet-Conv is weak to capture long-range contextual information in both the spatial and spectral directions. Moreover, existing BS methods fail to simultaneously consider global interaction between the spectral information and spatial information of different bands in a nonlinear fashion. In view of this, we propose to exploit the dual-attention mechanism, originally introduced in [19] for scene segmentation, to capture the long-range nonlinear contextual information in both the spectral and spatial directions.

Manuscript received May 13, 2020; revised July 10, 2020; accepted July 26, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61702065 and in part by the Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2018jcyjAX0033. *(Swalpa Kumar Roy and Sayantan Das contributed equally to this work.) (Corresponding author: Tiecheng Song.)*

Swalpa Kumar Roy is with the Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri 735102, India (e-mail: swalpa@cse.jgec.ac.in).

Sayantan Das is with the Computer Science and Engineering, Institute of Engineering and Management, Kolkata 700091, India (e-mail: sdas.codes@gmail.com).

Tiecheng Song is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: songtc@cqupt.edu.cn).

Bhabatosh Chanda is with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: chanda@isical.ac.in).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.3013235

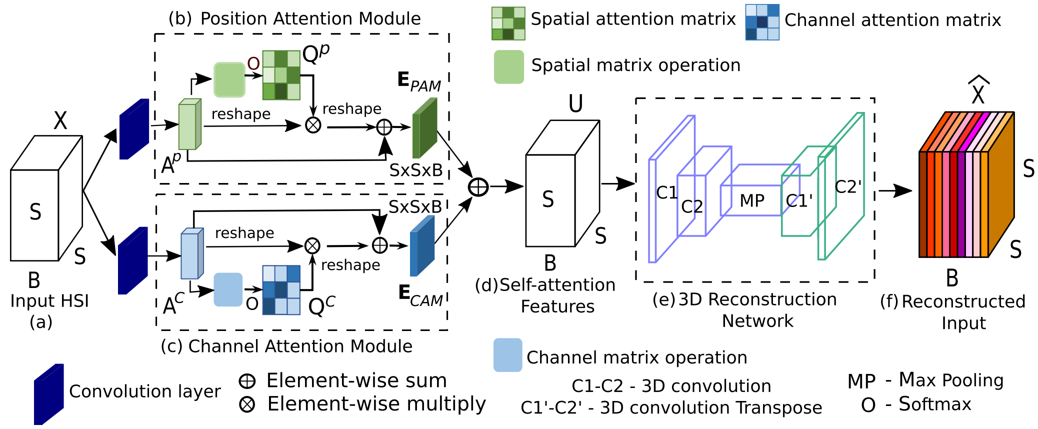


Fig. 1. Overview of the proposed dual-attention BS reconstruction network `DARecNet-BS`. (a) HSI. (b) PAM. (c) CAM. (d) Self-attention feature maps. (e) 3-D reconstruction network. (f) Restored input.

The contributions of this letter are highlighted as follows.

1) We propose a dual-attention reconstruction network for BS `DARecNet-BS`, an end-to-end unsupervised dual-attention reconstruction network for the BS task in the context of HSI domain.
2) The proposed network, which combines a position attention module (PAM) and a CAM, is coupled with a 3-D reconstruct network to capture long-range contextual information in both the spatial and spectral directions.
3) We demonstrate that our network can achieve state-of-the-art classification performance on several benchmark data sets.

II. Proposed DARecNet-BS Network

In this section, we present the `DARecNet-BS` for hyperspectral BS. `DARecNet-BS` employs a dual network containing a PAM and a CAM, followed by a 3-D reconstruct network to capture long-range contextual information in both the spectral and spatial directions (see Fig. 1). Before that, we first introduce some definitions and notations.

A spectral–spatial HSI that contains two spatial dimensions, namely the width W and the height H, and one spectral dimension B can be defined as $\mathbf{X_{orig}} \in \mathcal{R}^{W\times H\times B}$. All the pixels are classified into L_c land-cover classes denoted by $Y = (y_1, y_2, \ldots, y_{L_c})$. The pixel $\mathbf{x}_{i,j} \in \mathbf{X_{orig}}$, where $i = 1, \ldots, W$ and $j = 1, \ldots, H$, and hence, we can define the land-cover pixels as a spectral vector $\mathbf{x}_{i,j} = [x_{i,j,1}, \ldots, x_{i,j,B}] \in \mathcal{R}^B$. In the preprocessing step, neighboring regions of size $S \times S$ are extracted centered at pixel (i, j) from the original HSI data $\mathbf{X_{orig}}$. Depending on the neighboring region and spectral information, $\mathbf{x}_{i,j} \in \mathcal{R}^{B\times n\times n}$ can be further categorized into three sets, i.e., the pixel vector $\mathbf{x}_{i,j} \in \mathcal{R}^B$, the spatial region $\mathbf{x}_{i,j} \in \mathcal{R}^{S\times S}$, or the spectral–spatial region $\mathbf{x}_{i,j} \in \mathcal{R}^{S\times S\times B}$. To increase the discriminative power of any underlying network, the spectral–spatial information is jointly used, and the extracted spectral–spatial cubes $\mathbf{x}_{i,j}$ are stacked into X.

A. PAM

Aiming to find a set of informative spectral bands that can represent the whole band spectrum effectively, the PAM [19] can be used to recalibrate the strength of different spatial positions of the input. PAM takes HSI cubes $X \in \mathcal{R}^{S\times S\times B}$ as input and produces an output of spatial attention map $E_{\text{PAM}} \in \mathcal{R}^{S\times S\times B}$

$$E_{\text{PAM}} = \text{AttMod}^p(X; \theta^p) \tag{1}$$

where θ^p represents the trainable parameters involved in the PAM. The details of PAM are given step by step as follows.

Initially, X is passed through a convolutional layer, producing three sets of new features, i.e., $\text{Conv2D}(X) = A^p = \{A_1^p, A_2^p, A_3^p\}$, where the dimensions of A_1^p, A_2^p are reduced by a reduction factor, say $r = 8$, and the shapes become $A_1^p, A_2^p \in \mathcal{R}^{S\times S\times B/r}$ and $A_3^p \in \mathcal{R}^{S\times S\times B}$. Then, the obtained feature maps A_1^p, A_2^p, and A_3^p are reshaped into $\mathcal{R}^{V\times B}$, where $V = S \times S$ represents the number of pixels in a single band. Then, a matrix multiplication is performed between the reshaped feature maps $A_1^p \in \mathcal{R}^{V\times B}$ and $A_2^p \in \mathcal{R}^{V\times B}$, and a transpose operation is performed on $A_2^p \in \mathcal{R}^{B\times V}$ to satisfy the multiplication constraint. To calculate the resultant spatial attention map $Q^p \in \mathcal{R}^{V\times V}$, the matrix is passed through a softmax layer as follows:

$$q_{ji}^p = \frac{\exp\left(A_{1,i}^p, A_{2,j}^p\right)}{\sum_{i,j=1}^{V} \exp\left(A_{1,i}^p, A_{2,j}^p\right)} \tag{2}$$

where q_{ij} evaluates the positional impact between the ith and jth spatial features, which leads to greater correlation between their similar representation. After that, a matrix multiplication is again performed between the transpose of Q^p and A_3^p matrix. Then, a multiplication operation is performed with a trainable scalar parameter α^p, which is initially set to zero and gradually learned while training to provide more importance to the spatial attention [19]. Finally, the elementwise addition operation is performed with the input X to obtain the final spatial attention map $E_{\text{PAM}} \in \mathcal{R}^{S\times S\times B}$. The attention feature map generated from the PAM, i.e., E_{PAM}, can be, therefore, formulated as follows:

$$E_{\text{PAM},j} = \alpha^p \sum_{i=1}^{V} \left(q_{ji}^p A_{3,i}^p\right) + X_j. \tag{3}$$

As can be seen from (3), E_{PAM} selectively aggregates positionwise weighted sum of the learned features across all the ith and jth locations of input X in a global context under the guidance of the spatial attention map. The details of PAM are shown in Fig. 1(b).

B. CAM

Unlike PAM, the CAM [19] is used to find the strength of different spectral bands by recalibrating of the input. As shown in Fig. 1(c), the channel attention map $E_{\text{CAM}} \in \mathcal{R}^{S\times S\times B}$ can be directly calculated from input image $X \in \mathcal{R}^{S\times S\times B}$

$$E_{\text{CAM}} = \text{AttMod}^c(X; \theta^c) \tag{4}$$

where θ^c represents the trainable parameters associated with CAM. The details of CAM are described step by step as follows.

Initially, the input is stacked into $A^c = \{A_1^c, A_2^c, A_3^c\}$, where $A_1^c \in \mathcal{R}^{S\times S\times B}$ and $A_2^c \in \mathcal{R}^{S\times S\times B}$ are reshaped into $\mathcal{R}^{V\times B}$ and a matrix multiplication is performed between A_1^c and transpose of A_2^c. Then, the result is passed through a softmax layer to obtain the channel attention map $Q^c \in \mathcal{R}^{B\times B}$

$$q_{ji}^c = \frac{\exp\left(A_{1,i}^c, A_{2,j}^c\right)}{\sum_{i,j=1}^{B} \exp\left(A_{1,i}^c, A_{2,j}^c\right)} \tag{5}$$

where q_{ji}^c evaluates the impact of the ith channel on the jth channel. Finally, we perform a matrix multiplication between Q^c and the transpose of A_3^c and reshape the result into $\mathcal{R}^{S\times S\times B}$. The channel attention map E_{CAM} is obtained as follows:

$$E_{\text{CAM},j} = \alpha^c \sum_{i=1}^{B} \left(q_{ji}^c A_{3,i}^c\right) + X_j \tag{6}$$

where α^c is a trainable scalar parameter that controls the importance of the channel attention map across the input feature map X (it is initially set to 0 and allowed to learn during training). The abovementioned formulation aggregates channelwise weighted sum of the learned features across all the ith and jth channels of input $X \in \mathcal{R}^{S\times S\times B}$ in the global context guided by the channel attention map.

In order to gain more attention to the long-range contextual information, the generated feature maps from the abovementioned two attention modules, i.e., E_{PAM} and E_{CAM}, are aggregated using an elementwise sum fusion ($\oplus$) to model the position-channel attention features [see Fig. 1(d)]. The result is called self-attention feature $U_{\text{SAF}} \in \mathcal{R}^{S\times S\times B}$, which is formulated as

$$U_{\text{SAF}} = E_{\text{PAM}} \oplus E_{\text{CAM}}. \tag{7}$$

The self-attention feature helps to boost the feature discrimination ability compared with the original HSI data. To avoid the feature discrepancy, we model it without the convolutional layer before the feature fusion, as done in [19].

C. 3-D Reconstruction Network

To show the feature generalization ability, the original spectral bands are restored from the self-attention feature maps using an adapted 3-D reconstruction network (RecNet) by

$$\widehat{X} = \mathcal{F}_{\text{RecNet}}(U_{\text{SAF}}; \theta_e) \tag{8}$$

where θ_e is trainable parameters in RecNet and $\widehat{X} \in \mathcal{R}^{S\times S\times B}$ is the reconstructed output for the given input $X \in \mathcal{R}^{S\times S\times B}$. RecNet consists of two Conv3D layers and one maxpool3D, followed by two DeConv3D layers. Each convolution block consists of {Conv3D $\Rightarrow$ BatchNorm3D $\Rightarrow$ PReLU}. Each deconvolution block consists of {DeConv3D $\Rightarrow$ BatchNorm3D $\Rightarrow$ PReLU}, and finally, a Conv3D layer with BatchNorm3D is applied into RecNet to remove feature discrepancy. The shape of the kernel $1 \times 3 \times 3$ with a stride of size 1 is used throughout the convolutional/deconvolutional block in the network. The reconstruction performance of the network is measured by

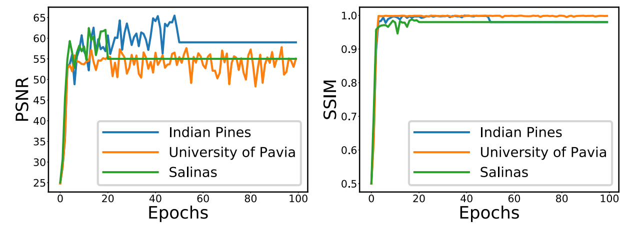


Fig. 2. PSNR and SSIM convergence curves for the reconstructed images in three benchmark data sets.

$$\mathcal{L}_1(\theta_b, \theta_e) = \frac{1}{2N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \|x_i - \hat{x}_i\|_1 \tag{9}$$

where $x \in X$, $\hat{x} \in \widehat{X}$, and N_{tr} is the number of training examples. The training is completed when the model converges or reaches the maximum iteration. The number of trainable parameters of the whole `DARecNet-BS` is about 200 976 on the Indian Pines (IP) data set.

To select the most informative spectral bands, the entropy is calculated from each band ($b_i \in B$) of the reconstructed output $\hat{X} \in \mathcal{R}^{S\times S\times B}$ using the following equation:

$$\mathcal{H}(b_i) = -\sum_h p(h)\log(p(h)), \quad \text{s.t.} \sum_h p(h) = 1 \tag{10}$$

where h is the gray level of histogram bins in a band consisting of $S \times S$ pixels, and $p(h) = (n(h)/(S \times S))$ is the probability that h occurs. Then, the entropy values are stored and sorted in descending order to select the top-k bands. According to Shannon's entropy theory, the larger the entropy is, the more information the bands will contain [9], [21].

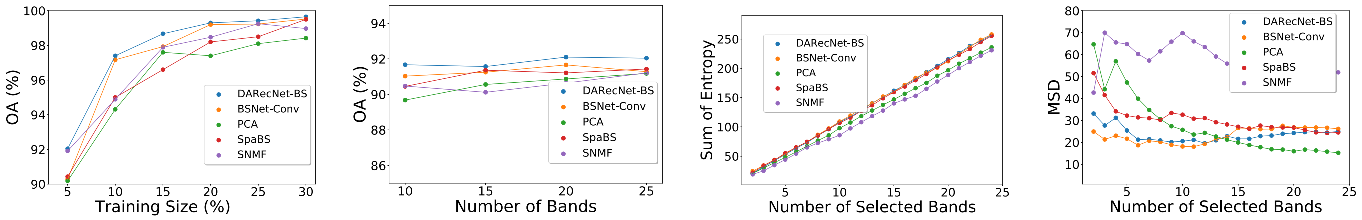
III. Experimental Results

Due to the nonavailability of proper ground truth, the efficiency of different BS methods is indirectly evaluated in terms of overall accuracy (OA), average accuracy (AA), statistical metric Kappa (κ), and some statistical analysis among the selected bands. The proposed `DARecNet-BS` is compared with well-known BS methods, such as SpaBS [12], principal component analysis (PCA) [5], SNMF [22], and BSNet-Conv [20]. To obtain robust classification performance, we use spectral–spatial residual network (SSRN) [16] in an end-to-end training fashion. The experiments are conducted using a 64-bit Ubuntu 18.04LTS operating system with NVIDIA Titan V 12-GB graphics processing unit. The whole framework is implemented in PyTorch with CUDA 10.1 enabled. We train `DARecNet-BS` by extracting 3-D patches of size $7 \times 7 \times B$, where band Bs from the IP, University of Pavia (UP), and Salinas (SA) data sets are set to 200, 103, and 204, respectively. Training is performed five times each using 200 epochs with a batch of size 32 on all the HSI data sets. The learning rate is set by cosine annealing scheduler, and the diffGrad [23] optimization is used for training.

A. Hyperspectral Data Sets

We use three well-known HSI data sets (i.e., IP, UP, and SA Scene)[1] to demonstrate the classification performance of the proposed `DARecNet-BS`.

[1] http://dase.grss-ieee.org/



(a) (b) (c) (d)

Fig. 3. (a) OA on varying training samples and (b) selected bands for the IP data set. The sum of entropy (c) and MSD (d) selected bands for the IP data set.

TABLE I

CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS USING SELECTED 25 BANDS FOR THE IP DATA SET WITH 5% TRAINING SIZE

No	SpaBS [12]	PCA [5]	SNMF [22]	BSNet-Conv [20]	DARecNet-BS
1	58.28 ± 42.02	32.50 ± 45.96	65.83 ± 46.56	**88.47 ± 5.14**	56.67 ± 40.67
2	85.57 ± 4.48	90.22 ± 4.60	91.29 ± 2.37	92.33 ± 1.76	**93.05 ± 5.33**
3	92.36 ± 2.47	91.44 ± 7.81	91.03 ± 1.37	95.48 ± 2.58	**95.56 ± 2.85**
4	85.04 ± 5.88	70.18 ± 2.40	87.51 ± 6.86	84.64 ± 7.29	**88.52 ± 2.38**
5	**98.78 ± 1.02**	89.00 ± 6.68	97.14 ± 0.95	77.87 ± 30.74	96.47 ± 4.98
6	98.39 ± 0.29	96.60 ± 0.75	99.14 ± 0.60	98.45 ± 1.37	**99.26 ± 1.92**
7	90.00 ± 14.14	61.90 ± 44.16	90.31 ± 7.58	**96.15± 5.43**	82.99 ± 8.88
8	97.91 ± 2.78	93.12 ± 4.91	95.30 ± 3.31	97.13 ± 2.92	**97.65 ± 3.99**
9	77.85 ± 19.01	89.85 ± 14.34	61.40 ± 43.89	94.73 ± 7.44	**96.73 ± 5.40**
10	**90.30 ± 3.06**	88.29± 3.83	88.56 ± 5.72	87.59± 3.65	85.66 ± 9.29
11	86.41 ± 2.10	89.64 ± 5.17	88.52 ± 5.99	91.30 ± 6.80	**93.67 ± 2.02**
12	84.57 ± 2.74	87.17 ± 9.25	**95.72 ± 1.53**	94.10± 2.31	81.98 ± 1.64
13	95.01 ± 5.63	99.29 ± 0.50	99.29 ± 1.00	99.46 ± 0.75	**99.79 ± 0.50**
14	95.72 ± 0.47	94.81 ± 1.50	94.56 ± 1.58	94.41 ± 2.39	**95.81 ± 1.27**
15	92.45 ± 1.52	89.07 ± 3.04	90.70 ± 5.40	**94.05 ± 6.39**	88.49 ± 3.94
16	96.82 ± 0.93	93.38 ± 3.34	94.95 ± 2.19	92.68 ± 4.80	**97.20 ± 1.49**
OA(%)	90.43 ± 0.81	90.18 ± 2.03	90.91 ± 2.99	90.28 ± 3.61	**92.04 ± 2.21**
AA(%)	89.09 ± 5.16	84.78 ± 5.91	89.45 ± 7.12	**92.48 ± 2.49**	89.42 ± 3.03
Kappa	0.890 ± 0.00	0.887 ± 0.02	0.907 ± 0.03	0.889 ± 4.04	**0.909 ± 0.02**

TABLE II

CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS USING 15 AND 20 BANDS FOR THE UP AND SA DATA SETS WITH 5% TRAINING SIZE

Data	Measure	SpaBS [12]	PCA [5]	SNMF [22]	BSNet-Conv [20]	DARecNet-BS
UP	OA(%)	97.84 ± 0.78	98.16 ± 0.47	98.46 ± 0.90	97.48 ± 1.21	**99.29 ± 0.32**
	AA(%)	98.07 ± 0.77	98.12 ± 0.76	97.93 ± 0.91	98.75 ± 0.64	**99.06 ± 0.25**
	Kappa	0.971 ± 0.01	0.975 ± 0.01	0.979 ± 0.01	0.972 ± 0.01	**0.990 ± 0.00**
SA	OA(%)	96.90 ± 0.70	90.50 ± 1.07	97.16 ± 1.31	97.48 ± 1.21	**97.99 ± 1.96**
	AA(%)	98.54 ± 0.30	93.59 ± 1.19	98.62 ± 0.48	98.65 ± 0.64	**98.74 ± 0.46**
	Kappa	0.965 ± 0.00	0.894 ± 0.01	0.968 ± 0.01	0.972 ± 0.01	**0.981 ± 0.52**

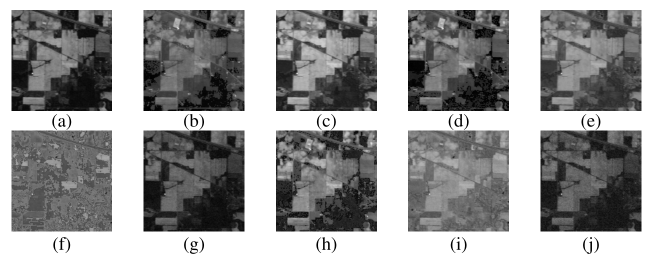


Fig. 4. Selected bands for IP data set with (a)–(e) high entropy and (f)–(j) low entropy, respectively.

The images in the IP data set contain 224 spectral bands in the wavelength range of 400–2500 nm with a spatial dimension of 145 × 145 pixels. The water absorption-based 24 spectral bands are not considered. The final IP data set is provided with the labeled classes for 16 types of vegetation.

The images in the UP data set contain 103 spectral bands in the wavelength range of 430–860 nm with a spatial dimension of 610 × 340 pixels. This data set is provided with the labeled classes for nine types of urban land covers.

The images in the SA scene data set contain 224 spectral bands in the wavelength range of 360–2500 nm with a spatial dimension of 512 × 217 pixels. The water absorption-based 20 spectral bands are not considered. The final SA data set is provided with the labeled classes for 16 types of fruits and vegetables.

B. Results on HSI Data Sets

To analyze the convergence of our proposed BS method, we perform BS from the reconstructed images after training using 100 epochs. The reconstruction quality always depends upon the computed structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). Fig. 2 shows SNR and SSIM convergence curves. As can be seen, the PSNR value of reconstructed images stabilizes to around 60 dB after 50 epochs for IP, and similarly, the SSIM value stays close to one after around 45 epochs. The large PSNR or the large SSIM value measures the quantitative quality of the reconstructed image generated from the network.

Table I shows the performance measure indices, i.e., OA, AA, and Kappa along with classwise accuracies computed under the subset of 25 bands with limited training samples of 5% for the IP data set. Table II shows the results of OA, AA, and Kappa for the UP and SA data sets using 5% training samples with 15 and 20 selected bands. One can see that the proposed `DARecNet-BS` method provides comparable or better classification performance over all data sets with a small standard deviation. Moreover, Fig. 3(a) and (b) shows the classification performance (OA) with respect to varying training samples and selected bands, respectively, for the IP data set. We see that `DARecNet-BS` achieves superior performance in terms of OA using most of the training sizes and the number of selected bands. Table III shows the performance gain (%) of OA, AA and Kappa for `DARecNet-BS` over models of PAM, CAM, and no attention mechanism. One can see that `DARecNet-BS` generally performs better than the models with attention PAM, CAM, and without any attention on all three data sets. It is also noted that the model with attention PAM or CAM can produce pretty good AAs only on the IP data set.

To analyze the redundancy among the selected top-k bands, we calculate an information theory-based criterion, i.e., mean spectral divergence (MSD) [24] that is expressed as

$$\mathrm{MSD} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} D_{\mathrm{KLS}}(b_i \| b_j) \tag{11}$$

where $b_i, b_j \subseteq B$, $D_{\mathrm{KLS}}(b_i \| b_j)$ is symmetric KL divergence given as $D_{\mathrm{KLS}}(b_i \| b_j) = D_{\mathrm{KL}}(b_i \| b_j) + D_{\mathrm{KL}}(b_j \| b_i)$, and $D_{\mathrm{KL}}(b_i \| b_j)$ is calculated from gray-level histogram bins. It can be inferred from (11) that the larger the value of MSD is, the less redundant information the selected bands contain. Fig. 3(c) and (d) shows the sum of entropy and MSD on the selected bands for the IP data set. It is also observed that SNMF [22] provides better MSD among the BS methods but unable to achieve good classification performance. The

TABLE III

PERFORMANCE GAIN (%) OF OA, AA, AND KAPPA FOR DARecNet-BS OVER MODELS OF PAM, CAM, AND NO ATTENTION

	IP			UP			SA		
	PAM	CAM	No Attention	PAM	CAM	No Attention	PAM	CAM	No Attention
OA	0.38	0.96	1.89	0.51	0.47	0.47	1.87	0.45	7.44
AA	-1.08	-1.03	-0.07	0.48	0.61	0.56	0.40	-0.14	2.51
Kappa	0.000	0.000	-0.004	0.007	0.006	0.006	0.025	0.009	0.086

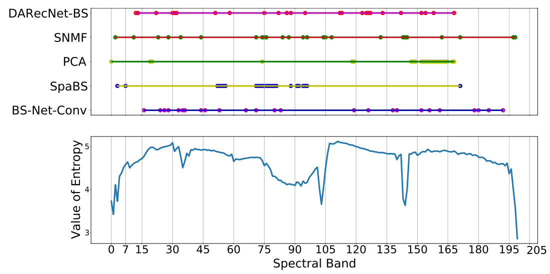


Fig. 5. (Top) Top-15 selected bands using different BS methods. (Bottom) Associated entropy value of each band.

selected top-five and bottom-five spectral bands for the IP data set are shown in Fig. 4. It is obvious that the top-five bands are more distinct due to large entropy than the bottom five. In addition, the top-15 selected bands using different BS methods and their entropy values are shown in Fig. 5. More detailed results can be found in the Supplementary Material.

IV. CONCLUSION

The letter introduces DARecNet-BS, an unsupervised dual-attention reconstruction network for hyperspectral BS. DARecNet-BS combines the position and spectral attention mechanisms to capture long-range contextual information in both spectral and spatial directions. Our network improves the feature representation ability for informative BS with less computational overhead. Experiments on three well-known data sets demonstrate superior performance using small training sets with less number of spectral bands.

ACKNOWLEDGMENT

The authors thank the Associate Editor and the anonymous reviewers for their outstanding comments and suggestions that greatly improve the quality of this work and Prof. Ananda Shankar Chowdhury from Jadavpur University, Kolkata, India, for the helpful discussion.

REFERENCES

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

[2] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (Overview and Toolbox)," 2020, *arXiv:2003.02822*. [Online]. Available: http://arxiv.org/abs/2003.02822

[3] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[4] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[5] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge-preserving features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7140–7151, Dec. 2017.

[6] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Noise-robust hyperspectral image classification via multi-scale total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1948–1962, Jun. 2019.

[7] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, pp. 304–315, Mar. 2014.

[8] M. Bevilacqua and Y. Berthoumieu, "Unsupervised hyperspectral band selection via multi-feature information-maximization clustering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 540–544.

[9] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 544–557, Jan. 2016.

[10] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, Mar. 2019.

[11] J.-M. Li and Y.-T. Qian, "Clustering-based hyperspectral band selection using sparse nonnegative matrix factorization," *J. Zhejiang Univ. Sci. C*, vol. 12, no. 7, pp. 542–549, Jul. 2011.

[12] K. Sun, X. Geng, and L. Ji, "A new sparsity-based band selection method for target detection of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 329–333, Feb. 2015.

[13] W. Zhang, X. Li, Y. Dou, and L. Zhao, "A geometry-based band selection approach for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4318–4333, Aug. 2018.

[14] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2019.

[15] S. K. Roy, S. Chatterjee, S. Bhattacharyya, B. B. Chaudhuri, and J. Platoš, "Lightweight spectral-spatial squeeze-and-excitation residual bag-of-features learning for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5277–5290, Aug. 2020.

[16] 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.

[17] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Multichannel pulse-coupled neural network-based hyperspectral image visualization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2444–2456, Apr. 2020.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[19] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[20] Y. Cai, X. Liu, and Z. Cai, "BS-nets: An end-to-end framework for band selection of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1969–1984, Mar. 2020.

[21] P. Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection," in *Proc. IEEE Workshop Adv. Techn. Anal. Remotely Sensed Data*, Oct. 2003, pp. 120–128.

[22] W. Sun, W. Li, J. Li, Y. M. Lai, "Band selection using sparse nonnegative matrix factorization with the thresholded earth's mover distance for hyperspectral imagery classification," *Earth Sci. Inform.*, vol. 8, pp. 907–918, Jan. 2015.

[23] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, "DiffGrad: An optimization method for convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 23, 2019, doi: 10.1109/TNNLS.2019.2955777.

[24] X. Geng, K. Sun, L. Ji, and Y. Zhao, "A fast volume-gradient-based band selection method for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7111–7119, Nov. 2014.