IEEE Transactions on Geoscience and Remote Sensing, Volume 60, June 2021, Article number 5506413
DOI: 10.1109/TGRS.2021.3075663

Novel Gumbel-Softmax Trick Enabled Concrete Autoencoder with Entropy Constraints for Unsupervised Hyperspectral Band Selection

He Sun, Jinchang Ren, *Senior Member, IEEE*, Huimin Zhao, Peter Yuen, and Julius Tschannerl

***Abstract*—As an important topic in hyperspectral image (HSI) analysis, band selection has attracted increasing attention in the last two decades for dimensionality reduction in HSI. With the great success of deep learning (DL)-based models recently, a robust unsupervised band selection (UBS) neural network is highly desired, particularly due to the lack of sufficient ground truth information to train the DL networks. Existing DL models for band selection either depend on the class label information or have unstable results via ranking the learned weights. To tackle these challenging issues, in this paper, we propose a gumbel-softmax (GS) trick enabled concrete autoencoder based UBS framework (CAE-UBS) for HSI, in which the learning process is featured by the introduced concrete random variables and the reconstruction loss. By searching from the generated potential band selection candidates from the concrete encoder, the optimal band subset can be selected based on an information entropy (IE) criterion. The idea of the CAE-UBS is quite straightforward, which does not rely on any complicated strategies or metrics. The robust performance on four publicly available datasets have validated the superiority of our CAE-UBS framework in classification of the HSIs.**

***Index Terms*—Unsupervised band selection; hyperspectral image; autoencoder; concrete random variable; information entropy.**

I. Introduction

As an emerging technology in the past few years, hyperspectral images (HSIs) have become increasingly popular on nondestructive inspection and characterization, owing to their rich spectral information spanning from visible to (near) infrared wavelengths. With the capability in identifying minor changes or differences of certain physical properties, such as moisture and temperature, and chemical components, HSIs have been successfully applied in a wide range of applications [1]–[4], especially in remote sensing, such as land cover analysis [5]–[7], precision agriculture [8], and object detection [9], [10], etc. Although the high-dimension spectral data is beneficial in discriminating different materials and objects, it has evitablly led to the 'Hughes phenomenon' [11], where the performance of the designed algorithms can be severely affected by insufficient training samples in comparison to the large number of spectral dimension. Moreover, the vast data volume of HSI also results in a huge computation cost, and the difficulty of data storage, transmission, and processing. Besides, the redundant information in HSIs may bring undesired properties and lower the efficiency of data analysis. Therefore, it is crucial to reduce the data dimension of the HSI data whilst preserving the essential discriminative information.

Although most of the feature extraction methods, such as the principal component analysis (PCA) [12], [13], the independent component analysis (ICA) [14], the wavelet transform [15], and the maximum noise reduction (MNF) [16] etc., can generate a discriminative and low dimensional feature set, the obtained features fail to preserve the physical characteristics of data acquired from the optical sensors. On the contrary, feature selection methods, which are also known as band selection, can choose a desired band subset and maintain the physical characteristics from the raw HSI [17]–[36].

According to whether the class label information is utilized, band selection methods can be grouped into three categories, i.e. supervised, semi-supervised and unsupervised. With the aid of the prior knowledge of the labeled pixels, (semi-)supervised methods select the optimal subset of bands by maximizing or minimizing a certain criterion [17], [20], [21]. However, they suffer from several intrinsical limitations. Firstly, it is impractical to collect sufficient training samples for each labeled category in real applications, especially for DL. Secondly, relying heavily on the classification performance can easily lead to overfitting and poor generalisity. Besides, the results can be of poor robustness as the selected band subset is subject to the randomly chosen training samples. As the label information is rarely available in real applications, unsupervised band selection (UBS) is focused in this paper.

Nowadays, DL-based methods have been successfully applied in many computer vision tasks and beyond [38]–[44], [46], [47]. In comparison to the conventional methods, DL-based approaches can automatically generate favourable features, not relying on manual intervention and subjective parameter settings. Many deep-learning models have already been applied in HSI, such as convolutional neural network (CNN) [39]–[41] and autoencoder (AE) [44], [46], [47], which are mainly for feature extraction and data classification [41], anomaly detection [46], [47], etc. Unlike the aforementioned tasks in HSI, there is no available ground truth in band selection to evaluate the chosen band subset for training the DL networks. Therefore, it is extremely challenging to determine

H. Sun is with the School of Computer Sciences, Beijing Institute of Technology, Beijing, China;
J. Ren is with National Subsea Centre, Robert Gordon University, Aberdeen, U.K. *Corresponding author: J. Ren (jinchang.ren@ieee.org)*;
H. Zhao is with School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou, China;
P. Yuen is with Electro-Optics & Remote Sensing, Centre for Electronics Warfare, Information & Cyber (CEWIC), Cranfield University, Swindon, U.K;
J. Tschannerl is with Apeel Sciences, California, U.S.
Manuscript submitted 2020.

a desired band subset by using a DL-based UBS method.

In this paper, we have proposed a novel AE-based DL framework for UBS in HSI. By training an AE with the defined reconstruction loss, the optimal band subset can be determined for reconstructing the original HSI cube. Different from our previous work in [44], the optimal band subset is obtained directly from the trained AE without the assistance of ranking the significance of each band. The major contribution of this paper can be highlighted as follows:

1) A concrete end-to-end AE-based UBS framework, CAE-UBS, is proposed, in which the optimal band subset with the desired number of bands can be easily determined according to the best reconstruction of the original HSI. Rather than using continuous real-numbers as the weights in the encoder module, a novel concrete layer is implemented with a binary weight of 1 and 0 to indicate whether the corresponding band is selectable or not. It is only because of the introduced Gumbel-Softmax, the obtained discrete weight matrix can be transformed to continuous variables for optimization of the selected band subset during the backpropagation. To the best of our knowledge, this is the first time to employ the Gumbel-Softmax trick to obtain the desired band subset directly in AE deep learning-based UBS in HSI.
2) Being implemented in an unsupervised manner, the proposed CAE-UBS network is found to be efficient and robust for UBS according to the reconstruction loss and the classification accuracy of the HSI. With the aid of an information entropy-based criterion, the desired band subset can be determined with much less computational cost than other DL methods.
3) In the proposed CAE-UBS framework, a weight matrix from a fully connected (FC) layer has been utilized to initialize the class probabilities, which can effectively improve the classification performance. The superior performance of our proposed CAE-UBS framework has been validated on four commonly used HSI datasets to demonstrate its merits over a number of state-of-the-art (SOTA) UBS and one supervised methods, especially a more robust performance with less trainable parameters and no label information needed.

The rest of this paper is organized as follows. Section II introduces the related UBS methods and AE-based DL methods. Section III details the proposed framework, including CAE-based band selection and optimal band subset searching. The experimental results on four HSI datasets are presented and discussed in Section IV. Finally, Section V concludes the paper along with some future directions.

II. Related work

In the last two decades, a number of UBS approaches have been proposed, which can be grouped into four main categories, i.e. ranking-based, clustering-based, searching-based methods, and sparsity-based. For each category, a detailed literature review is summarized below. In addition, the background information of the AE and AE-based UBS methods will also be introduced in this section.

In ranking-based band selection, many efforts have been made to evaluate and rank the importance of the raw spectral bands so as to determine the most significant bands from the raw spectral cube. In [22], a maximum-variance PCA (MVPCA) criterion was utilized to estimate the band prioritization. As MVPCA considers the representative and discriminative ability of each individual band but ignores the correlation between the chosen bands, the selected band subset is generally lack of robustness. In Chang and Wang [23], a constraint band correlation (CBS) strategy is proposed for ranking-based UBS. Four criteria are adopted in the CBS framework for choosing the highly correlated dependent bands, including the band correlation minimization (BCM), the band dependence minimization (BDM), the band correlation constraint (BCC), and the band dependence constraint (BDC). Although noisy band which has less correlation to all other bands will be discarded, similar to MVPCA, the band subset selected from CBS still contains a high degree of redundancy. For ranking-based methods, the result is usually quite redundant because of the high correlation between the selected bands, due mainly to focusing only on the performance of each band rather than the relationship between different bands.

Unlike ranking-based approaches, clustering-based methods firstly group all the bands into clusters before selecting the most representative band from each cluster. By clustering adajacent bands together under various similarity metrics, the correlation of the bands chosen from different clusters can be naturally reduced. In [24], a hierarchy clustering algorithm (WaLuDi/WaLuMi) is proposed based on the Ward's linkage, which clusters the bands by maximizing the inter-cluster variance whilst minimizing the intra-cluster variance. According to the Ward's linkage theory, the chosen band from each cluster is the most representative one hence the formed band subset will be robust. However, the WaLuDi/WaLuMi method suffers from a huge computational cost due to its hierarchy architecture.

Some researchers have dedicated their work to improving the clustering-based method by combining with some ranking strategies. Inspired by the fast density-peak-based clustering (FDPC) [45], an enhanced FDPC (E-FDPC) [26] was proposed to rank each band by considering the local density and the intra-cluster distance simultaneously, where the introduction of the intra-cluster distance has effectively reduced the correlation between the selected bands. In [27], Wang et al. have proposed an optimal clustering framework (OCF) for UBS in HSI with two objective functions, inspired by the top-ranked cut and the normalized cut for effective band clustering. Afterwards, three ranking strategies are utilized to rank the bands within each cluster for band selection, where the top-ranked band in each cluster is chosen to form the selected band subset. Although these clustering methods achieve a good performance, noisy bands are prone to become a single cluster and lower the robustness. To tackle this, an adaptive distance-based band hierarchy (ADBH) [28] has been proposed recently to reflect the hierarchy structure of HSI and produce any number of desired band subsets, whilst the effect of noisy bands can be suppressed. In clustering-based methods, choosing only the most representative band from each cluster

may be insufficient as the second representative band in one cluster may contain more information than the first one in another cluster. Thus it is more important to rank the band subset as a whole rather than individually, which can also avoid the effect of noisy bands, as it can easily form a separate cluster in such approaches.

With a given objective function and a search strategy, searching-based methods determine an optimal band subset by exploring different possible combinations of the bands. In [30], the Volume Gradient band selection (VGBS) method is introduced, where the defined 'volume' information can be obtained from the estimated covariance matrix of all bands. By assuming the most redundant band has the maximum gradient, the VGBS can iteratively remove redundant bands until the desired number of bands is reached. By developing a structure-aware metric for measuring band informativeness and independence, Zhu et al [34] proposed a dominant-set extraction UBS (DSEBS)method. As a greedy search-based method, DSEBS tackles the UBS as a clustering problem. As searching for the optimal subset is an NP-hard problem and too costly, the used meta-heuristic or evolutionary algorithms usually produces a suboptimal solution [34]. In [52], the relationship between each band and the entire hypercube is determined through the linear reconstruction, and a desired band subset can be searched by removing the effect of noisy bands, the proposed optimal neighborhood reconstruction (ONR) method has achieved a good performance on UBS.

Apart from the searching-based methods, the sparsity-based methods utilize the sparse representation (SR) to explore the underlying structures within the HSI data [32]. The multitask sparsity pursuit (MTSP) [31] searches the optimal band subset with the aid of the SR and the immune clonal strategy. Although in SR based methods it is quite straightforward to select the informative bands based on the estimated sparse coefficients, the overall computational complexity is still quite high especially in constructing the SR matrix for large-scale HSI datasets [27].

Recently, DL and its variations have shown great superiority in extracting more effective features in HSI. Cai et al [40] have proposed an end-to-end CNN-based model for band selection, where the final band subset is determined by ranking the average of the learned weights for each band. Unlike other deep learning-based neural networks, the basic idea of AE-based feature selection is to learn the hidden representations that can effectively reconstruct the input data. Due to its strong ability to explore both linear and nonlinear structures among the extracted features, AE has been successfully applied for feature selection in high dimensional data in an unsupervised manner [44]. For UBS in HSI, the AE-based methods are not as popularly used as the aforementioned other categories of the methods. In our previous work [44], the input weights of the AE are utilized to select the most significant bands in an unsupervised way. However, there are several drawbacks for this kind of methods. The generated representation from the encoder is more like a combination of the raw data, where the weight values of nodes in the encoder layer can be both positive and negative. Some bands are chosen only because they have large absolute weights, which does not fully represent their significance. Besides, the aforementioned methods rely on the ranking value or the weight to choose the desired band, which can inevitably suffer from the disadvantages of ranking-based UBS methods, especially the high redundancy between the chosen bands. These will be tackled in our proposed approach as detailed in the next section.

III. Methodology

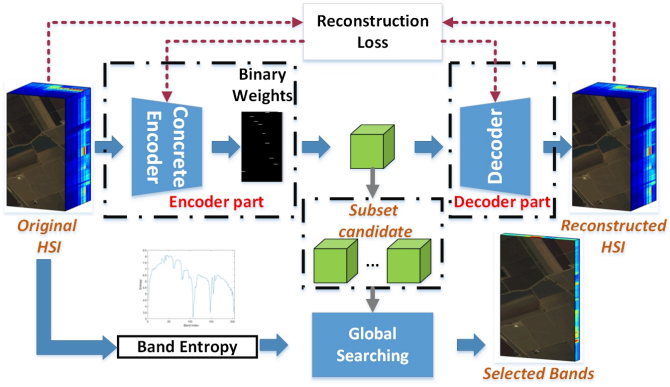


Fig. 1: The flowchart of the proposed CAE-UBS framework.

In this section, our proposed CAE-UBS framework will be presented in detail, including the concept of CAE-based band selection, determing the optimal band subset, and computational complexity analysis. According to the flowchart shown in Fig. 1, first a HSI hypercube is taken as the input to the designed CAE. Potential band subsets can be acquired based on minimizing the reconstruction error of the hypercube with the designed CAE. After calculating the IE of each candidate of band subsets, the band subset with the maximum IE will be chosen as the result of band selection. Relevant details are presented as follow.

A. CAE based band selection

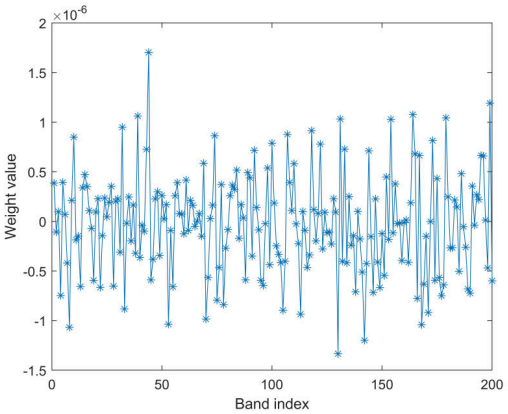


Fig. 2: Weight values of one column in the learned weight matrix W^{en}, the horizontal and vertical axes represent the band index and weight values of Indian Pine dataset, respectively.

In general, a standard AE includes one encoder and one decoder module. The encoder represents the mapping between the input data and the hidden representation while the decoder is to reconstruct the input data from the hidden representation. Let us project an HSI image into a matrix $X = [X_1, \dots X_i, \dots, X_m] \in R^{D \times m}$ denote the projected data from a hypercube, where m represents the total number of samples in the HSI image and D is the number of spectral bands. Based on that, the encoder function can be depicted as $H_i = \sigma_{en}(X_i W^{en} + b^{en})$ and the decoder function that reconstructs the input data as $\hat{X}_i = \sigma_{de}(H_i W^{de} + b^{de})$, where the H_i is the hidden representation of the input data and the $\hat{X}_i$ is the reconstructed data. σ_{en} and σ_{de} are the activation functions, and W and b are the weighted matrices and bias vector of each module, respectively. For the UBS work, the w_d^{en} within the input weight matrix $W^{en} = (w_1^{en}, ..., w_d^{en}, ..., w_D^{en})$ actually measures the dth band and represents the contribution of the dth band in the reconstruction process. The AE can be trained with the supervision of the reconstruction loss:

$$L = \frac{1}{2m} ||X - \hat{X}||_F \tag{1}$$

In our previous work [44] and other similar work [40], the desired band subset can be chosen by ranking the learned weight W^{en} from the encoder part. The basic assumption here is that a highly ranked weight indicates more important of the corresponding band. However, the weight learned from AE in general cannot represent the significance of each band. For example, Fig. 2 shows the learned input weight with one column in the learned weights matrix W^{en} of the Indian Pines dataset. Although the positive values represent the contribution of this band, it has several negative values. Besides, the motivation of AE-based band selection is to select the most significant bands for spectrum reconstruction, yet the input weight based band selection strategy seems not linked to this objective. Therefore, it is inappropriate to choose the band according to the weight values.

As the purpose of the AE-based band selection is to learn an important hidden representation from the input data for HSI reconstruction, it would be more reasonable to extract the desired band subset from the encoder part as the key latent features of the raw data. Inspired by this, we aim to determine a sparse input weight matrix, whose values can be only 1 and 0, indicating the corresponding band is selected or not. In this manner, the weight of the bands that do not contribute to the reconstruction will be 0, otherwise will be 1. Moreover, the extracted band subset will be optimal as the weights of the chosen bands are jointly learned. However, this sparse weight matrix cannot be updated during the backpropagation in a standard AE as each column of this matrix is a one-hot vector, i,e, a non-differentiable discrete variable. To tackle this problem, we have introduced a novel concrete AE for the UBS, where the sparse matrix can be estimated with the aid of concrete distribution [48], [49] as detailed in next subsection.

In our proposed CAE-UBS framework, we have employed the above concrete random variables to select the input bands. Let the desired number of bands in the band subset be k, a new

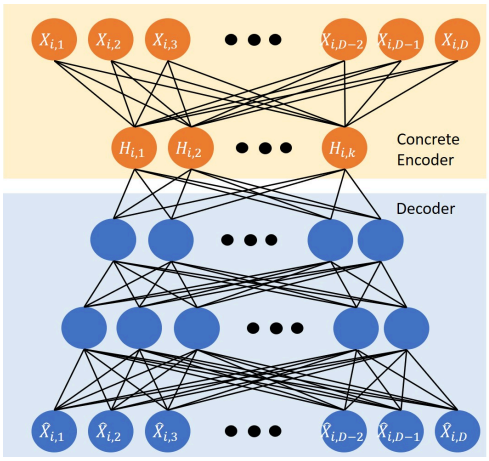


Fig. 3: The diagram of the designed concrete autoencoder, where the $X_{i,D}$ represents the Dth band of the original HSI data X_i and $\hat{X}_{i,D}$ is its corresponding reconstructed value, the H_i denotes the chosen band subset with k bands

sparse weight matrix S will be built with a size of $D \times k$. For each column of the weight matrix S, a D-dimensional concrete random variable S_k is sampled following (3). In this way, the output of the encoder module is $H_i = X_i S$ for an input sample X_i. As S_k is a one-hot vector, it can select a band to reconstruct the original data. Thus, the composed weight matrix S becomes a desired sparse matrix, in which the selected k bands can be directly identified without introducing another criterion. With the aid of the introduced concrete random variable and reparameterization, the forward propagation can generate a band subset, and the backpropagation will optimize the band selection result.

B. Concrete distribution

The GS distribution, also referred as the concrete distribution, is defined to produce a continuous distribution over a discrete variable, e.g., a one-hot vector. As a reparameterization trick, the Gumble-softmax trick can efficiently sample z, i.e. a one-hot vector, from a catergorical distribution with class probabilities α_k, where g_k is the sample drawn from Gumbel $(0, 1)^1$ and k is the element-wise index of the generated one-hot vector z.

$$z = one\_hot \left\{ \arg\max_k [g_k + log(\alpha_k)] \right\} \tag{2}$$

As the above operation is non-differentiable, which cannot be back-propagated in the network for optimization. To tackle this issue, the GS distribution [48] using the softmax function is proposed as a continuous differentiable approximation to replace the $\arg\max$ function in (2) for calculating the continuous relaxation of the one-hot vector z, where the k^{th} element of the generated sample S from the concrete distribution is given by:

$$S_k = \frac{exp((g_k + log(\alpha_k))/T)}{\sum_{d=1}^{D} exp((g_k + log(\alpha_k))/T)} \tag{3}$$

the temperature parameter T controls the relaxation of the one-hot vector, where S_k will nearly equal to 1 when T approaches to 0. With the reparameterization trick, S_k becomes differentiable when estimating the gradient in the process of backpropagation.

C. Optimal band subset searching

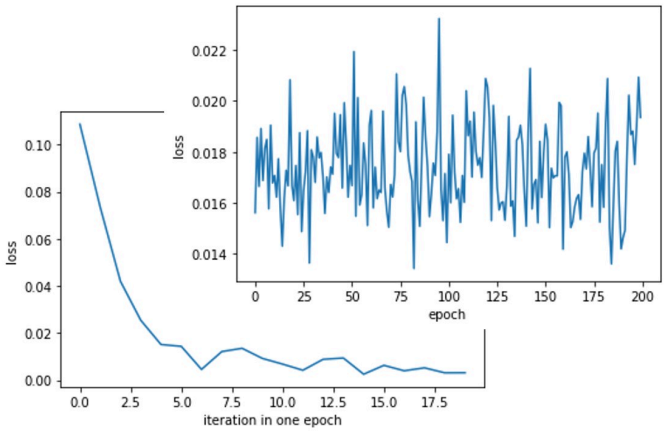


Fig. 4: Top-right: Training loss curve in 200 training epochs on the Indian Pines dataset. Bottom-left: The training loss from the 100th training epoch on the Indian Pines dataset, where the number of iterations equals to the number of batches.

For searching the desired band subset efficiently, we randomly divide all samples from an hypercube into different batches in a similar way as other DL models [38]. In this way, multiple band subsets can be obtained during each epoch. Let N be the number of band subsets determined in one epoch, it actually equals to the number of iterations, i.e. the number of batches, in each epoch. Although a band subset is selected according to the minimized reconstruction error, it can be potentially only the local optimal solution due to the random selection of the batch, where searching for a global optimal band subset is still needed. To this end, a simple yet robust IE-based searching strategy is introduced in our CAE-UBS framework as detailed below [35].

Generally, there are several motivations for considering the global searching strategy. The first is to find an efficient way to determine the optimal band subset whilst avoiding a huge computational cost. Nowadays, most of the efficient UBS methods are still not DL-related, where in AE-based UBS the optimal band subset is assumed to be the one with the best reconstruction ability. We have further speculated that the desired band groups contain more information than other subsets, which is beneficial for spectrum reconstruction. To this end, we have defined a global searching strategy using information theory [35], the Shannon IE, where the IE of band X_i is defined as:

$$IE(X_i) = -\int_{X_i} P(X_i) log(P(X_i)) dx \tag{4}$$

where $P(X_i)$ denotes the probability density function of X_i, which can be usually estimated by [27], [35]. Based on the determined IE for each band, the band subset with the largest accumulated IE is chosen as the desired band subset, and the result is considered as the global optimal solution [27], [35]. As this search strategy is quite straightforward and efficient, it has been adopted in the proposed CAE-UBS approach.

D. New weights initialization for improved efficiency

To further improve the efficiency of DL-based UBS in HSI, a rapid convergence of the designed network is essential for significantly reducing the computational complexity. In existing GS-based methods [48], [49], the class probability α_k is often randomly initialized in small positive values for exploring different linear combinations of the inputs, which may affect the convergence of the network and the result of band selection. In our CAE-UBS framework, we initialize the α_k with the weight matrix from a FC layer to regularize the learning process, where the initialized weight matrix has the same size of the composed weight matrix S. In this way, α_k are initialized within $(-\sqrt{D}, \sqrt{D})$, adaptive to the number of bands, which is further normalized to $(0, 1)$ to indicate the class probability. The efficacy of the proposed initialization has been further validated in the comparison experiments in the next section.

To obtain the desired band subsets without too much computational cost, another key point is the efficiency in generating potential candidates. As one training epoch can produce N candidates, this will end up with a large search space after a few epochs. Besides, more training epochs increase the running time of the whole framework. To find the optimal band subset efficiently, we need to reduce the number of training epochs. With our proposed CAE, we have found that the convergence is faster due to the data volume as the HSI data is around 100 thousands pixels about several hundreds MB but RGB dataset is usually GB level. In Fig. 4, the training loss, i.e. the reconstruction loss, of 200 training epochs on the Indian Pines dataset is presented. As seen, the training loss is obviously reduced in each epoch in Fig. 4. Based on that, we conclude that the proposed network can converge within only one epoch, and the optimal band subset can be chosen from the generated N candidates. In this manner, the efficiency of the proposed CAE-UBS framework can be ensured.

E. Merits of CAE-UBS

With the concrete random variable-based AE and IE based searching strategy, our CAE-UBS framework can determine an optimal band subset for the effective reconstruction of the original spectral data. Different from other AE-based band selection frameworks, we have formulated the band selection task as a searching-based task by maximizing the accumulated IE of the desired band group instead of ranking the significance of each band. Moreover, the proposed CAE can solve the problem of backpropagation even with a discrete variable in the UBS task, which enables the designed network able to be trained with the reconstruction loss L. Being trained in a self-learning way without introducing any class label information, the proposed CAE-UBS has the potential to inspire more related research on the DL-based band selection in the future.

Algorithm 1 CAE-UBS

Input: Raw HSI data $X = [X_1, \dots X_i, \dots, X_m] \in R^{m \times D}$, desired number of bands k.
Initialize: Hyperparameters Initialization :Adam optimizer with learning rate lr, Temperature parameter T, Batch size B.
BEGIN
Estimate IE of each band in X
while the first epoch **do**
    Encoder module: Initialize α_k;
    Encoder module: learn S based on (3);
    $H_i = X_i S$;
    Save N band subsets
    Decoder module;
    Update reconstruction loss L based on (1);
    Backpropagation with optimizer;
end while
Global optimal band subset searching with IE (4) of each band and N band subsets;
Output: Band subset n.
END

The whole process of the proposed CAE-UBS is summarized in Algorithm 1, where the performance is further discussed in the next section.

IV. Experimental results

Due to the lack of the ground truth in the UBS task, the performance of band selection is usually indirectly assessed by evaluating the classification accuracy with the selected bands. In our experiments, the proposed CAE-UBS is compared with several SOTA methods based on the classification performance as detailed below.

A. Datasets

Four commonly used HSI remote sensing datasets are used in our experiments. The first is the Indian Pines dataset, which was captured by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the North-Western Indian, USA in 1992. The raw data has 224 spectral bands with the wavelength ranging from 0.4-2.5μm. It has a spatial size of 145×145 pixels, in which 10249 pixels are manually labelled in 16 land-cover categories. Often, the dataset is corrected to have 200 bands after the removal of 24 noisy and water absorption bands.

The second is the Pavia University (PaviaU) dataset, which was collected by the Reflective Optics System Imaging Spectrometer (ROSIS) system over the Engineering School of the University of Pavia, Italy. The commonly used PaviaU dataset is a cropped version, which consists of 610×340 pixels with a spectral range of 0.43-0.86 μm. This dataset has 42776 pixels labelled in 9 land-cover classes.

The third is the Salinas dataset, which was also acquired by the AVIRIS over the Salinas Valley, California, USA in 1998. Therefore, it shares the same wavelength range with the Indian Pines dataset in 224 spectral bands. The spatial size is 512×217, in which 54129 pixels are labelled in 16 classes. After removing the noisy and water absorption bands, 204 bands are remained for experiments.

The last is the Botswana dataset, which was captured by NASA EO-1 satellite sensor over OKAvango Delta, Botswana in 2011. The original dataset contain 242 bands ranging from 400-2500nm. With a spatial size 1476×256 pixels, in total 3248 pixels are labelled in 16 semantic classes. After the removal of 97 noise-corrupted bands, a corrected dataset with 145 bands is often utilized.

B. Settings

For quantitative evaluation of the classification results with the selected bands as features, three commonly used metrics derived from the confusion matrix are adopted, including the overall accuracy (OA), the average accuracy (AA), and the Kappa coefficient. OA represents the percentage of corrected classified pixels, and AA is the mean classification accuracy over all classes. The Kappa coefficient is introduced to estimate the reliability of the obtained results.

For performance evaluation, we have compared our method with a few SOTA UBS algorithms, including the optimal clustering framework (OCF) (TRC-OC-EFDPC) [27], the band selection with dominant set extraction (DSEBS) [34], the volume gradient band selection (VGBS) [30], WaLuDi/WaLuMi [24], the enhanced fast-peak-based clustering (E-FDPC) [26], the Adaptive Subspace Partition Strategy (ASPS) [29], and the Adaptive Distance based Band Hierarchy (ADBH) [28]. These SOTA UBS algorithms are introduced as follows:

1) OCF [27]: a SOTA clustering-based method with a leading performance in the UBS of HSI.
2) DSEBS [34]: one of the most representative searching-based UBS methods. By developing a structure-aware measurement for band informativeness and independence, it tackles the UBS as a greedy-searching problem, which has achieved a relatively good performance on several public datasets.
3) VGBS [30]: also a searching-based method, frequently cited in UBS [27], [28]
4) WaLuDi/WaLuMi [24]: Although being proposed earlier than other compared methods, they are still classical clustering-based methods and frequently cited in many literatures [27]–[29], [34].
5) E-FDPC [26]: Different from other ranking-based methods, an enhanced fast density-peak-based clustering proposed to rank each band by considering the local density and the intra-cluster distance simultaneously, which has a leading performance in ranking-based methods.
6) ASPS [29]: a novel clustering-based method with a robust performance in the UBS of HSI.
7) ADBH [28]: an adaptive distance-based band hierarchy based UBS to reflect the hierarchy structure of HSI for easily producing any number of desired band subsets whilst suppressing the effect of noisy bands.

TABLE I: Classification results for the Indian Pines dataset using the raw data or selected bands (averaged on 3-30 bands).

Classifier	OCF	VGBS	DSEBS	WaLuDi	WaLuMi	E-FDPC	ADBH	ASPS	Ours	Raw data
OA by KNN(%)	64.52±3.45	59.06± 3.26	**68.74±3.39**	63.35±2.89	51.90±8.28	61.21±1.76	67.93±3.32	64.27±5.24	67.94±2.74	67.65±0.02
AA by KNN(%)	55.03±3.96	46.98±2.73	54.99±2.68	51.48±3.01	39.67±9.42	47.01±2.87	**57.76±3.78**	52.77±5.66	57.55±2.3	54.22±0.01
Kappa by KNN	0.59±0.04	0.53±0.04	**0.64±0.04**	0.58±0.04	0.45±0.1	0.55±0.03	0.63±0.04	0.59±0.06	**0.64±0.03**	0.62±0.01
OA by SVM(%)	75.39±6.21	66.66±5.51	74.34±5.6	73.99±4.03	65.89±12.63	69.52±5	**76.43±5.48**	73.29±7.75	75.98±5.4	79.33±0.01
AA by SVM(%)	73.36±9.02	62.2±7.56	71.89±8.83	72.33±5.72	57.84±21.51	65.76±10.57	**74.13±9.02**	70.66±12.03	74.12±6.96	71.47±0.01
Kappa by SVM	0.72±0.07	0.62±0.07	0.70±0.07	0.70±0.05	0.60±0.16	0.65±0.06	**0.73±0.08**	0.70±0.1	**0.73±0.07**	0.75±0.01

TABLE II: Classification results for the PaviaU dataset using the raw data or selected bands (averaged on 3-30 bands).

Classifier	OCF	VGBS	DSEBS	WaLuDi	WaLuMi	E-FDPC	ADBH	ASPS	Ours	Raw data
OA by KNN(%)	83.19±1.72	83.91±1.84	81.92±2.36	83.96±1.94	85.07±2.05	84.24±0.92	83.18±1.86	**85.69±1.68**	85.34±2.4	85.73±0.02
AA by KNN(%)	79.12±2.03	80.38±2.29	76.38±3.2	79.86±2.27	81.88±2.38	80.68±1.56	78.76±2.86	**82.45±1.92**	81.18±3.43	82.02±0.01
Kappa by KNN	0.77±0.02	0.78±0.03	0.75±0.03	0.78±0.03	0.80±0.03	0.79±0.01	0.77±0.03	**0.81±0.02**	0.80±0.03	0.81±0.01
OA by SVM(%)	88.4±3.42	88.47±4.28	87.52±4.07	89±3.33	89.15±3.02	87.06±2.02	88.69±3.57	83.49±3.73	**89.92±4.01**	91.64±0.01
AA by SVM(%)	86.11±4.95	84.93±7.91	84.55±5.68	86±5.75	**86.32±4.75**	83.97±3.68	85.61±6.34	77.29±3.65	86.24±6.47	88.12±0.01
Kappa by SVM	0.85±0.05	0.85±0.07	0.83±0.05	0.85±0.05	0.86±0.05	0.83±0.03	0.86±0.06	0.78±0.03	**0.87±0.00**	0.89±0.00

For a fair comparison, the original codes from the authors and the default parameters are used. Besides, the classification results from the original data are also included (shown as 'Raw data' in this paper).

The proposed CAE-UBS method also has several parameters. In the training process, we have employed the Adam optimizer with a learning rate of 1e-3, where the training epoch is set to 1 for efficiency. In DL, a large batch size can improve the training efficiency than a small one, yet it may suffer from poor convergence and poor generalization. As a result, a proper batch size needs to be determined, which is suggested to be linked to the size of the image [38]. In our experiments, the batch sizes for Indian Pines, PaviaU, Salinas, and Botswana datasets are empirically set to 512, 8192, 8192, and 8192 by considering their spatial sizes, i.e. the number of pixels. These parameters are found to produce particular good results in band selection in our proposed approach. In addition, the activation function of the designed stacked decoder is ReLU. For the temperature parameter, we follow the schedule in [49].

For the classification part, two commonly used classifiers, K-Nearest Neighbourhood (KNN) [50] and Support Vector Machine (SVM) [11], are employed with the selected band subsets from each method as features. In our experiments, the parameters of KNN and the SVM are optimized through a 10-folds cross-validation. We use 10% of the randomly chosen labelled samples as the training set, and the rest for testing. For the compared methods, the experiments are repeated 10 times, and the average metrics are reported. As our approach is DL-based, the chosen band subset can be affected by some stochastic issues. Therefore, the output band subset slightly different in each run of experiments.

Nowadays, DL-based methods usually report their best results from the trained models in other computer vision tasks such as image segmentation and object detection [38]. Considering that non-deep-learning based conventional approaches may produce fixed results, it is unfair they are compared with the best results from DL approaches. Therefore, we randomly choose five groups of the band selection results from our CAE-UBS framework, where the selected bands are taken as features for classification in 10 repeated runs. Afterwards, the average metrics of these five subsets in 50 total runs are reported for comparison with the peers.

For the hardware and software settings, the proposed CAE-UBS framework is implemented on the Pytorch 1.1.0 package without CUDA. All other band selection methods and the classification part are implemented on the MATLAB 2019a. All experiments are done with an Intel i5-8400 CPU, 16GB rem, with the results reported below.

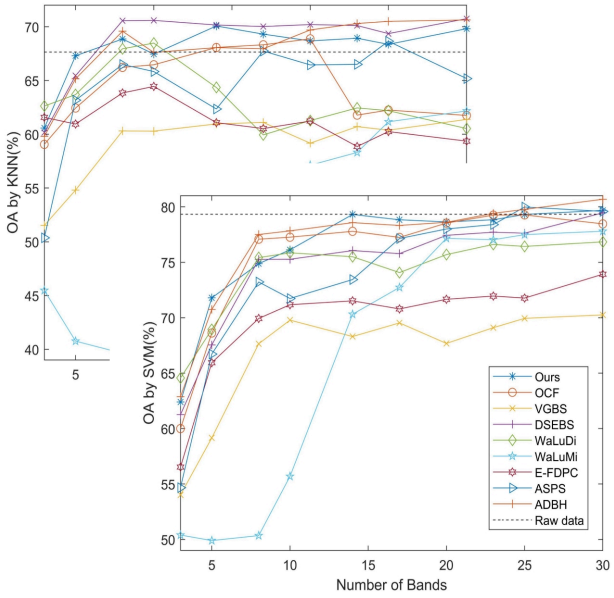


Fig. 5: OA curves on the Indian Pines dataset with different UBS methods. Bottom-Left: OA by KNN; Top-right: OA by SVM.

C. Results and discussions

For performance assessment, the OA curves of all methods on the four HSI datasets are generated 3-30 chosen bands

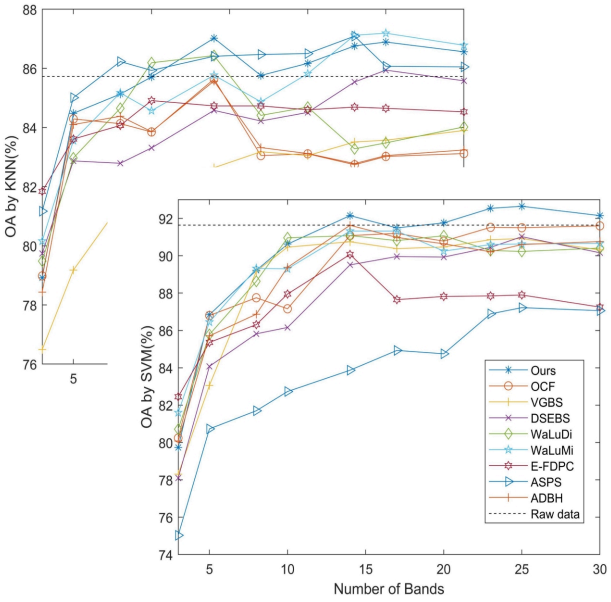


Fig. 6: OA curves on the PaviaU dataset with different UBS methods. Bottom-Left: OA by KNN; Top-right: OA by SVM.

and shown in Figs. 5-8. As seen, most of the comparing methods compete with the performance using raw data when the number of chosen bands is around 30. Besides, we have compared the average OA, AA, and the Kappa coefficient and their corresponding standard variation under various number of selected bands. A detailed comparison of each method on the four datasets is given in Tables I-IV, respectively, where the best result is highlighted in bold except those from the raw data.

To summarize the experimental results from the four datasets, some extended discussions are given below. In particular, we will discuss in three aspects, i.e. the performance of our method, the comparison between our method and BS-Net, another SOTA DL-based UBS method, and analysis of the computational time of each method.

1) Comparison results in different datasets: For the Indian Pines dataset, the classification results from different approaches are presented in Fig. 5 and Table I. As seen in Fig. 5, our method has a robust performance on both KNN and SVM classifiers. Although the performance is the second best on the KNN classifier, the difference to the first place, the DSEBS, is marginal. When more than 20 bands are selected, only the DSEBS, ADBH and our CAE-UBS approaches performance well. For the SVM classifier, our results are also quite stable, especially when the number of selected bands is beyond 20. Although our approach does not outperform others in all cases, a robust OA curve has validated its superiority. Table I shows the classification results of all methods. As seen, the proposed method along with the ADBH and DSEBS methods have better performance than the rest on the KNN classifier. However, the performance of DSEBS with on SVM seems not as good as on the KNN classifier. For our approach, it has achieved the second best results with both classifiers, indicating its robustness.

For the PaviaU dataset, the results are compared in Fig. 6 and Table II, and again our proposed method has shown quite stable performance. For the KNN classifier, our approach has an increasing OA. Although the WaLuMi method produces the best results with the KNN classifier when more than 25 bands are selected, it does not perform well with a small number of selected bands, and the performance with the SVM seems not robust as shown in Fig. 6. With the SVM classifier, our method has achieved a more robust OA than all others. Considering both the classifiers, our generated OA curves are steadier, which has validated the robustness of our CAE-UBS method. This has been further verified in the quantitative results in Table II, where our approach has achieved the best OA on SVM and the second best OA on the KNN classifier. Although the ASPS has achieved the best classification result on the KNN, its performance on the SVM is rather poor. For the OCF and ADBH, their performance on the PaviaU dataset are not good.

Fig. 7 and Table III show the classification results for the Salinas dataset. In Fig. 7, our method again has achieved nearly the best performance on the KNN classifier and a robust performance on the SVM. Although our approach is not the best on the KNN when less than 15 bands are chosen, its superiority accelerates when more bands are selected. Although OCF has a better performance when less than 15 bands are selected, its OA curve on the KNN classifier is not as stable as ours. For the VGBS and the WaLuDi methods, they fail to produce satisfying results on the KNN. For the SVM classifier, most of the methods have achieved a robust performance except for the WaLuDi, whilst our is the third best slightly behind the ADBH and OCF methods. This is also consistent with the results in Table III, where our approach is the thrid best on the KNN and the SVM, whilst the difference between ours and the two leading ones are minor.

For the Botswana dataset, the classification results from different UBS approaches are presented in Fig. 8 and Table IV. As seen in Fig. 8, our method has the most robust OA curves than all others on both classifiers. Although the WaLuDi has a better performance when 5 or less bands are chosen, our CAE-UBS method has a more stable OA curve. With the SVM classifier, the WaLuDi does not perform well when more than 5 bands are chosen. The VGBS has a poor performance with less selected bands even though it has the best result when 30 bands are chosen. As shown in Table IV, our approach has the best average OA on the SVM classifier. For the KNN classifier, we have the second best averaged OA with a marginal difference to the WaLuDi, which demonstrates again its robustness.

2) Further Result Analysis: Although the proposed method has obtained quite good results with the two popular classifiers on the four HSI datasets, the OA is not always the best which can be explained as follows. In fact, the network architecture and the strategy for searching the optimal band subset used in the proposed method are relatively simple. Taking the proposed CAE-UBS framework as a baseline, its performance can be further improved by introducing a larger neural network or certain regularization terms such as spatial constraints.

TABLE III: Classification results for the Salinas dataset using the raw data or selected bands (averaged on 3-30 bands).

Classifier	OCF	VGBS	DSEBS	WaLuDi	WaLuMi	E-FDPC	ADBH	ASPS	Ours	Raw data
OA by KNN(%)	88.33±0.73	84.84±1.39	87.28±2.37	85.65±1.02	86.81±2.1	87.43±1.43	**88.48±0.65**	86.58±1.73	88.18±1.34	87.70±0.01
AA by KNN(%)	**93.32±0.64**	88.94±2.15	92.36±1.87	91.32±1.45	91.24±2.79	92.61±1.49	93.3±0.75	91.82±1.69	92.95±1.22	93.27±0.01
Kappa by KNN	0.87±0.01	0.83±0.02	0.86±0.03	0.84±0.01	0.85±0.02	0.86±0.01	**0.87±0.01**	0.85±0.01	**0.87±0.01**	0.86±0.01
OA by SVM(%)	92.22±1.71	91.66±1.92	90.87±3.46	90.14±2.15	91.46±3.21	91.82±1.59	**92.45±1.61**	90.84±3.27	91.95±2.04	92.87±0.00
AA by SVM(%)	95.8±1.26	95.18±1.64	94.91±2.58	94.27±2.38	94.53±3.82	95.38±1.55	**95.85±1.24**	94.63±2.69	95.51±1.6	96.42±0.00
Kappa by SVM	0.91±0.02	0.91±0.02	0.90±0.04	0.89±0.02	0.90±0.04	0.91±0.02	**0.92±0.02**	0.90±0.04	**0.92±0.00**	0.92±0.01

TABLE IV: Classification results for the Botswana dataset using the raw data or selected bands (averaged on 3-30 bands).

Classifier	OCF	VGBS	DSEBS	WaLuDi	WaLuMi	E-FDPC	ADBH	ASPS	Ours	Raw data
OA by KNN(%)	80.53±2.95	78.21±2.9	79.99±3.71	**82.36±0.66**	80.76±2.9	79.6±4.13	81.15±1.66	77.52±7.83	82.25±2.85	82.44±0.01
AA by KNN(%)	77.81±3.06	75.22±3.15	77.39±3.76	**79.72±0.75**	78.16±2.74	76.98±4.13	78.52±1.69	74.95±7.77	79.21±3.79	80.11±0.01
Kappa by KNN	0.79±0.03	0.76±0.03	0.78±0.04	**0.81±0.01**	0.79±0.03	0.78±0.05	0.8±0.02	0.76±0.09	**0.81±0.03**	0.81±0.02
OA by SVM(%)	86.37±3.47	86.43±4.39	85.11±4.19	87.72±1.11	86.97±3.84	85.35±4.17	86.41±2.94	83.06±7.83	**88.1±2.96**	89.94±0.01
AA by SVM(%)	87.21±3.59	87.26±4.73	86.14±4.23	88.6±1.05	87.94±3.82	86.15±4.33	87.45±2.89	84.04±7.88	**88.82±3.44**	91.54±0.02
Kappa by SVM	0.85±0.04	0.85±0.05	0.84±0.05	**0.87±0.01**	0.86±0.04	0.84±0.04	0.85±0.03	0.82±0.08	**0.87±0.03**	0.89±0.01

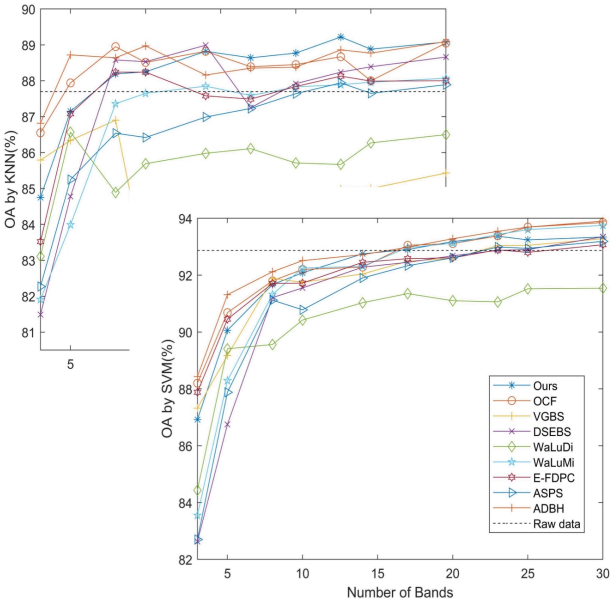


Fig. 7: OA curves on the Salinas dataset with different UBS methods. Bottom-Left: OA by KNN; Top-right: OA by SVM.

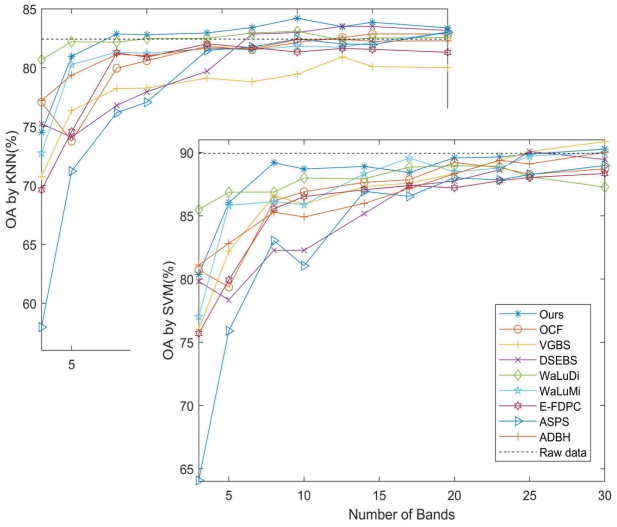


Fig. 8: OA curves on the Botswana dataset with different UBS methods. Bottom-Left: OA by KNN; Top-right: OA by SVM.

Actually, the quite satisfactory results on four datasets from three different sensors, i.e. the AVIRIS, ROSIS, and the NASA EO-1 sensors, have validated the robust performance and high generalization capability of the proposed network. To this end, it is safely to say that the proposed method can generate a global optimal solution in most cases.

As shown in the previous subsection, our proposed CAE-UBS framework can usually produce better results when more bands are selected. For example, our OA curve in Fig. 6 outperforms all others when more than 15 bands are chosen. As our method is searching-based, a larger search space with more bands tends to produce better results. Therefore, it is prone to find the optimal band subset from the increased number of band combinations, which validates the searching ability of our developed DL-based UBS method.

3) Comparison with other deep learning-based UBS methods: To further evaluate the effectiveness of the proposed method, we have compared it with one SOTA DL-based UBS method, the BS-Net [40], and the AE-UBS [44]. For BS-Net, the indexes of selected bands provided by the authors are used to test the classification accuracy. For the three test datasets, Indian Pines, PaviaU, and Salinas, the numbers of selected bands given in [40] are 25, 15, and 20, respectively. As a result, we compare our approach with BS-Net using the same number of selected bands. The selected bands by BS-Net and our method are listed in the Appendix, where the BS-Net has two groups of results, i.e. by using FC networks (BS-Net-FC) and convolutional neural networks (BS-Net-Conv) for evaluation and comparison. In addition, we have listed five groups of results from our approach and one group of results from our previous approach [44]. Taking the selected bands

TABLE V: Comparison results between other deep learning-based UBS methods [40], [44] and the proposed method on the first three datasets.

Dataset	Classifier	CAE-UBS	AE-UBS [44]	BS-Net-FC [40]	BS-Net-Conv [40]
Indian Pines	OA by KNN(%)	68.36±0.01	68.07±0.01	64.76±0.00	**71.91±0.02**
	AA by KNN(%)	58.77±0.01	51.70±0.01	53.18±0.01	**61.97±0.02**
	Kappa by KNN	0.65±0.01	0.64±0.01	0.59±0.01	**0.68±0.03**
	OA by SVM(%)	79.31±0.01	77.99±0.01	76.85±0.01	**80.66±0.00**
	AA by SVM(%)	79.62±0.011	76.18±0.01	73.96±0.00	**80.39±0.00**
	Kappa by SVM	**0.78±0.01**1	0.75±0.01	0.74±0.01	**0.78±0.01**
PaviaU	OA by KNN(%)	85.66±0.00	84.70±0.01	**87.11±0.01**	83.99±0.01
	AA by KNN(%)	82.31±0.00	81.04±0.00	**84.38±0.01**	79.92±0.01
	Kappa by KNN	0.81±0.01	0.79±0.01	**0.83±0.01**	0.78±0.01
	OA by SVM(%)	**92.84±0.01**	85.00±0.01	92.63±0.01	92.75±0.01
	AA by SVM(%)	**90.93±0.01**	75.79±0.01	90.75±0.01	90.59±0.00
	Kappa by SVM	**0.91±0.01**	0.79±0.00	0.90±0.00	0.90±0.01
Salinas	OA by KNN(%)	**88.77±0.01**	88.49±0.01	88.18±0.01	87.11±0.01
	AA by KNN(%)	**93.70±0.00**	93.16±0.01	93.45±0.01	92.87±0.01
	Kappa by KNN	**0.87±0.01**	**0.87±0.01**	**0.87±0.01**	0.86±0.00
	OA by SVM(%)	**93.18±0.01**	92.80±0.00	92.70±0.00	91.99±0.01
	AA by SVM(%)	**96.35±0.01**	96.20±0.01	96.23±0.01	95.96±0.00
	Kappa by SVM	**0.92±0.01**	**0.92±0.01**	**0.92±0.00**	0.91±0.01

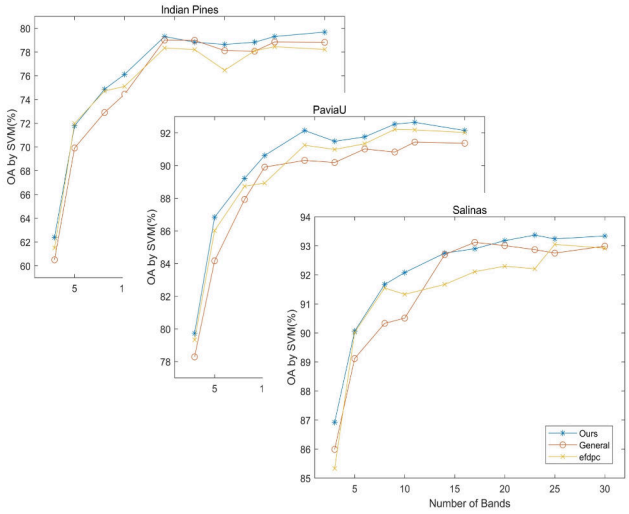


Fig. 9: OA curves on the three datasets with different methods, 'Ours' represents the proposed method, 'General' is the proposed method with general GS initialization, and the 'efdpc' denotes the proposed method with the E-FDPC ranking instead of IE. Bottom-left: Indian Pines; Mid: PaviaU, Top-right: Salinas.

as the spectral features, we can then use the classification results as an indicator to evaluate the efficacy of the band selection approaches. In Table V, quantitative results in terms of OA, AA and Kappa from the BS-Net, AE-UBS, and CAE-UBS are given for comparison, using the KNN and SVM classifiers on the three datasets. As seen in Table V, the BS-Net-Conv has the best performance on the Indian Pines dataset with both classifiers, followed slightly behind, especially for SVM, is our CAE-UBS method. Nevertheless, our method significantly outperforms the BS-Net-FC with both classifiers. For the PaviaU dataset, the proposed approach has the best performance with the SVM classifier and the second best performance with the KNN classifier, while the BS-Net-FC has achieved the best performance with the KNN classifier. Surprisingly, BS-Net-Conv has produced much worse results than the BS-Net-FC, especially for the KNN classifier, although it has the best results on the Indian Pines dataset. This has indicated relevant lack of robustness of the BS-Net model in different datasets. Finally, for the Salinas dataset, our CAE-UBS method has yielded the best performance with both classifiers, though it seems BS-Net-FC performs slightly better than BS-Net-Conv. Besides, our proposed CAE-UBS method has outperformed the AE-UBS with both classifiers in the three validated datasets.

It is worth noting that the reported band subsets chosen from the BS-Net are selective the best to produce the highest classification accuracies. For our approach, we have used the averaged classification results from five randomly chosen band subsets. To this end, the superior performance has validated the robustness and efficacy of the proposed CAE-UBS method. As our method is implemented using a less complicated network with only the FC layers, the performance could be further improved by introducing the convolutional kernels or adding more layers, which will be explored in the future.

4) Effect of α_k initialization: Generally, the GS distribution initializes the α_k with small positive values. In our CAE-UBS, we assume the general GS initialization method cannot reflect the class probabilities, we have employed the weight from a FC layer to initialize the α_k. To illustrate the effectiveness of our proposed initialization approach, we have compared the classification results with the general GS initialization methods and ours. The classification results in terms of OA with the SVM classifier on the first three datasets are shown in Fig. 9. As seen, our initialization method has produced a more robust OA curve than the general one, especially in the PaviaU dataset. Accordingly, it can consistently produce improved classification accuracy under the same number of selected bands, which has validated the superiority of the proposed initialization scheme.

5) Analysis of the IE: To further analyse the effect of the utilized (IE) criterion in our proposed CAE-UBS approach, we have replaced it by E-FDPC [26], a popularly used method to rank the band importance [27], [28]. The ranking values obtained by E-FDPC are employed to determine the desired band subset, and the results with the SVM classifier are also compared in Fig. 9. As seen in Fig. 9, the proposed CAE-UBS with the IE criterion has consistently outperformed the variation with the E-FDPC for band ranking rather than the IE, especially on the Salinas dataset. The robust performance here has validated the superiority of the IE criterion used in the proposed CAE-UBS approach.

TABLE VI: Computational time(s) of different UBS methods with 30 selected bands vs. the average OA in 3-30 selected bands with the SVM classifier on the four datasets.

Methods	Avg. OA(%)	Indian Pines	Pavia U	Salinas	Botswana
Ours	**86.49**	0.75	1.9	1.5	4.7
ADBH	85.99	0.35	2.69	2.23	7.73
OCF	85.59	0.7	0.65	1.13	2.5
WaLuDi	85.21	41.95	99.7	198.51	322.25
DSEBS	84.46	**0.2**	1.02	1.05	8.63
E-FDPC	83.44	0.97	6.85	3.11	22.35
WaLuMi	83.37	14.04	13.82	29.68	77.59
VGBS	83.31	0.54	**0.24**	**0.82**	**0.47**
ASPS	82.67	0.56	1.89	1.01	7.72

6) Comparison of computational time: To evaluate the efficiency of the proposed approach, we have also compared in Table VI the computational times of various methods with 30 selected bands. Meanwhile, the average OA from the SVM classifier in 3-30 selected bands for all the four datasets is used to indicate the efficacy of these band selection algorithms. As seen in Table VI, our method has outperformed all others yet with a comparable computational complexity to the conventional methods without DL. Although OCF seems quite efficient, its OA is about 0.9% lower than ours. For WaLuDi and WaLuMi, their computational complexity is quite high due to the time-consuming process in calculating the mutual information. As it only requires one training epoch, the proposed CAE-UBS approach has actually provided an efficient and effective solution for UBS in HSI.

As an indicator of the computational complexity of the DL-based approaches, the numbers of parameters of our proposed CAE-UBS approach, and BS-Net [40] are compared in Table VII. As seen, our CAE-UBS has much less trainable parameters than the BS-Net and AE-UBS. However, the classification accuracies are comparable to or even superior than BS-Net as shown in Table V. Note that the reported computational time including the training process for HSI reconstruction is implemented on CPU, hence the efficiency can be further improved with the aid of GPU like other DL-based band selection approaches such as the BS-Nets [40]. In comparison to BS-Nets implemented on a 11GB GPU, our CAE-UBS algorithm implemented on a CPU is about 1000 times faster, yet the classification results are very comparable or superior. Thanks to the GS trick and entropy constraints, this has validated again the great potential of CAE based UBS in HSI.

TABLE VII: Number of parameters in CAE-UBS.

No.of.Parms	Indian Pines	PaviaU	Salinas	Botswana
CAE-UBS	43244	17489	51860	29669
AE-UBS	85533	25437	85760	46283
BS-Net-FC	152592	-	-	-
BS-Net-Conv	590288	-	-	-

V. Conclusions

Although a few unsupervised approaches have been proposed for hyperspectral band selection in the last two decades, the results in general show lack of robustness due to the bank ranking schemed adopted, whilst the DL-based approaches often suffer from huge computational burden due to numerous training epochs requested. In this paper, we have proposed a novel CAE-UBS framework for unsupervised hyperspectral band selection. With the introduced CAE, the collaborative behaviour of the bands during the HSI reconstruction process can be exploited for searching the candidates of potential band subsets. By implementing a novel encoder layer with the GS trick, a discrete matrix can be generated to choose the desired band subset, where parameters of the proposed CAE can be learned by the constraints of the reconstruction loss. In addition, maximizing the accumulated IE is found to be an effective global searching strategy to determine the optimal band subset. As the proposed CAE can produce satisfactory results with only one training epochs, its computational time has been significantly reduced to the same level as conventional methods. The robust performance from experiments on four publicly available datasets has fully demonstrated the efficiency and efficacy of the proposed CAE-UBS framework.

Although the proposed approach produces overall the best performance, the results vary in the four datasets. In the future, we will focus the development of a multi-task network for selecting more discriminative bands for classification, aiming to achieve more consistent performance in different datasets. In addition, we will explore other band selection applications in HSI beyond image classification, such as spectral unmixing and HSI reconstruction.

Acknowledgment

The authors would also like to thank the authors of OCF, VGBS, DSEBS, WaLuDi/WaLuMi, E-FDPC, ASPS for providing the original codes. Besides, we also would like to thank the authors of [51] for providing their band selection results.

Appendix

To further demonstrate the performance of our proposed method, we have compared our method with a novel supervised deep learning-based method, the BHCNN [51]. Due to the limited computational resource, we could not run the code of [51] directly. Instead, we have compared our method with it using the 30 selected bands in the Indian Pines and PaviaU datasets, based on the band selection results provided by [51]. Taking these selected bands as input features, we can produce the classification results using KNN and SVM for comparison. As seen in Table A-I, the propsoed method outperforms the BHCNN [51] in the selected 30 bands.

The results of selected bands by the proposed method and the BS-Net on three HSI datasets, Indian Pines, PaviaU and Salinas, are shown in Tabel A-II, where the number of bands are the same as used in BS-Net, i.e. 25, 15 and 20 for the three datasets. For visual comparison of the importance of each band, we compute the band significance according to the chosen times of each band divided in five repeated experiments. The band significance from these three datasets are shown in Fig. A. 1.

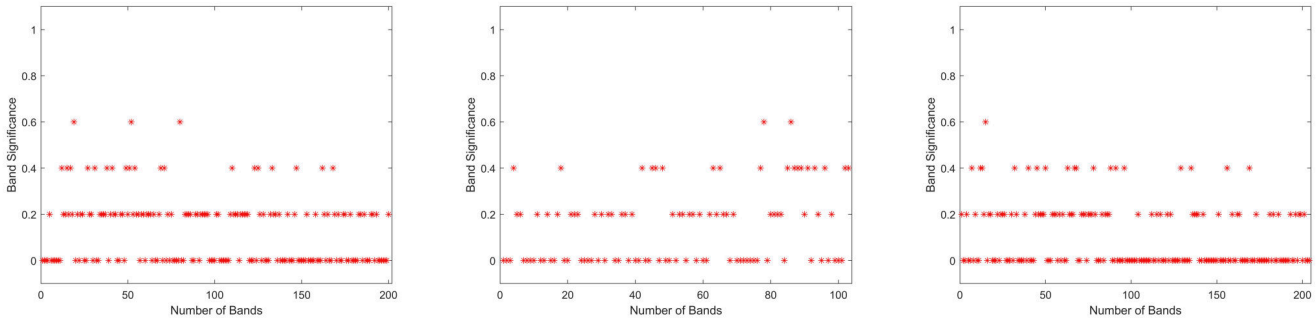
References

[1] J. Tschannerl, J. Ren, F. Jack, et al., "Potential of UV and SWIR hyperspectral imaging for determination of levels of phenolic flavour compounds in peated barley malt," *Food. Chem.*, vol. 270, pp. 105-112, Jan. 2019.

[2] J. Tschannerl, J. Ren, H. Zhao, et al., "Hyperspectral image reconstruction using Multi-colour and Time-multiplexed LED illumination," *Opt. Laser. Eng.*, vol 121, pp. 352-357, Oct. 2019.

[3] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE Trans. Cybern.*, vol.47, no. 4, pp. 884-897, Apr. 2017.

[4] T. Qiao, J. Ren, C. Craigie, J. Zabalza, C. Maltin, and S. Marshall, "Singular spectrum analysis for improving hyperspectral imaging based beef eating quality evaluation", *Comput. Electron. Arg.*, vol. 115, pp. 21-25, Jul. 2015.



(a) (b) (c)

Fig. A.1: The band significance based on the given band subset of CAE-UBS in the Table A-II, where the horizontal axis and vertical axis denote the band index and the probability of each band to be selected for Indian Pines (a), PaviaU (b), Salinas (c), respectively.

TABLE A-I: Comparison results between a supervised deep learning-based BS methods, BHCNN [51] and the proposed method on the first two datasets.

Dataset	Classifiers	CAE-UBS	BHCNN [51]
Indian Pines	OA by KNN(%)	**69.82±0.01**	66.09±0.01
	AA by KNN(%)	**54.36±0.01**	51.99±0.01
	Kappa by KNN	**0.64±0.01**	0.61±0.01
	OA by SVM(%)	**79.68±0.01**	78.14±0.01
	AA by SVM(%)	**77.08±0.01**	74.05±0.01
	Kappa by SVM	**0.77±0.01**	0.75±0.01
PaviaU	OA by KNN(%)	**86.56±0.00**	85.32±0.02
	AA by KNN(%)	**83.44±0.01**	81.72±0.01
	Kappa by KNN	**0.82±0.01**	0.80±0.01
	OA by SVM(%)	**92.15±0.01**	89.53±0.01
	AA by SVM(%)	**90.15±0.01**	83.57±0.01
	Kappa by SVM	**0.90±0.01**	0.86±0.01

[5] J. Zabalza, J. Ren, J. Zheng, et al., "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging", *Neurocomputing*, Vol. 185, pp. 1-10, 2016.

[6] L. Zhang, Q. Zhang, B. Du, et al., "Simultaneous spectral-spatial feature selection and extraction for Hyperspectral Images," *IEEE Trans. Cybern.*, vol.48, no. 1, pp. 16-28, Jan. 2018.

[7] Z. Feng, S. Yang, M. Wang, and L. Jiao, "Learning dual geometric low-rank structure for semisupervised hyperspectral image classification," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2018.2883472.

[8] J. Zabalza, J. Ren, Z. Wang, et al., "Singular spectrum analysis for effective feature extraction in hyperspectral imaging," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1886-1890, Nov. 2014.

[9] J. Zabalza, J. Ren, J. Zheng, et al., "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in Hyperspectral Imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4418-4433, Aug. 2015.

[10] C. Zhao, X. Li, J. Ren, and S. Marshall, "Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery", *Int J Remote Sens.*, vol. 34 no. 24, pp. 8669-8684, Dec. 2013.

[11] M. Pal and G. M. Foody, "Feature Selection for Classification of Hyperspectral Data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297-2307, May 2010.

[12] J. Zabalza, J. Ren, M. Yang, et al., "Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing," *ISPRS J. Photogramm. Remote Sens.,* vol. 93, pp. 112-122, Jul. 2014.

[13] 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.

[14] J. Wang and C.-I Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analsysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586-1600, Jun. 2006.

[15] T. Qiao, J. Ren, Z. Wang, et al.. "Effective denoising and classification of hyperspectral images using curvelet transform and singular spectrum analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 119-133, Jan. 2017.

[16] A. A. Green, M. Berman, P. Switzer and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65-74, Jan. 1988.

[17] W. Sun, G. Yang, J. Peng and Q. Du, "Hyperspectral band selection using weighted kernel regularization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3665-3676, Sept. 2019.

[18] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.,* vol. 8, no. 1, pp. 138-142, Jan. 2011.

[19] X. Cao, T. Xiong, and L. Jiao,"Supervised band selection using local spatial information for hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 329-333, Mar. 2016.

[20] S. Patra, P. Modi, and L. Bruzzone, "Hyperspectral band selection based on rough set," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5495-5503, Oct. 2015.

[21] H. Su, H. Yang, Q. Du, Y. Sheng, "Semisupervised band clustering for dimensionality reduction of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 6, pp. 1135-1139, Nov. 2011.

[22] C. I. Chang, Q. Du, T.-L. Sun, and M. L. G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631-2641, Nov. 1999.

[23] C. I. Chang, and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575-1585, Jun. 2006.

[24] A. Martínez-Usómartinez-Uso, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158-4171, Dec. 2007.

[25] Y. Yuan, J. Lin, and Q. Wang, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1431-1445, Mar. 2016.

[26] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 88-102, Jan. 2016.

[27] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910-5922, Oct. 2018.

[28] H. Sun, J. Ren, H. Zhao, et al., "Adaptive distance-Based band hierarchy (ADBH) for effective hyperspectral band selection," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2020.2977750.

[29] Q. Wang, Q. Li, and X. Li, "Hyperspectral band selection via adaptive subspace partition strategy," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published. doi:10.1109/JSTARS.2019.2941454.

TABLE A-II: Selected bands by the proposed method and BS-Net on three datasets.

Dataset	Methods	Selected bands
Indian Pines	CAE-UBS	12 13 15 17 35 38 40 49 51 52 73 86 92 102 112 125 126 133 137 146 153 158 175 190 200
		5 12 17 19 27 36 37 41 65 68 84 90 109 117 119 123 128 132 133 134 147 168 174 178 186
		19 23 27 31 41 43 50 52 53 54 62 80 83 85 89 93 94 110 111 115 147 162 165 169 183
		14 15 16 18 21 24 28 29 31 34 38 47 49 51 59 69 71 80 118 142 162 163 168 191 193
		19 42 46 52 54 55 56 58 61 63 69 71 75 80 95 96 101 110 113 116 123 125 135 171 184
	AE-UBS	25 31 32 39 41 54 55 62 69 91 92 101 106 128 135 140 149 150 151 159 180 186 188 190 196
	BS-Net-FC	1 6 12 13 23 27 35 36 39 47 52 53 61 66 72 75 76 89 95 101 161 165 166 180 185
	BS-Net-Conv	17 25 25 29 34 36 37 45 47 54 67 72 81 84 120 127 139 141 153 157 162 179 181 186 193
PaviaU	CAE-UBS	4 18 23 37 45 46 48 56 64 77 82 87 89 93 94
		17 18 30 36 48 51 62 67 78 86 88 89 96 98 103
		5 11 32 42 46 53 54 63 66 81 83 85 91 93 102
		6 21 22 39 42 45 63 65 77 78 85 86 91 102 103
		4 14 28 33 57 59 65 69 78 80 86 87 88 90 96
	AE-UBS	2 5 12 14 15 18 24 28 33 37 42 62 68 91 102
	BS-Net-FC	18 21 39 63 66 67 75 79 80 82 86 91 93 96 99
	BS-Net-Conv	4 5 8 17 32 39 43 49 54 72 79 81 91 99 100
Salinas	CAE-UBS	7 9 12 32 38 45 48 60 63 67 68 76 78 79 123 151 169 186 193 201
		1 14 15 17 25 30 32 40 55 64 65 72 78 87 115 135 139 156 181 196
		4 13 15 18 26 33 44 45 50 58 66 86 91 96 104 121 136 137 162 169
		7 13 24 40 47 50 56 67 73 77 83 88 91 117 129 142 163 173 198 199
		12 15 22 46 49 54 63 68 71 75 85 88 96 112 129 135 138 156 159 183
	AE-UBS	10 27 31 32 43 44 51 81 89 119 120 123 124 165 171 173 179 193 196 202
	BS-Net-FC	6 9 17 19 22 38 41 44 46 50 54 55 57 62 78 159 177 180 190 198
	BS-Net-Conv	19 20 47 92 96 98 105 109 117 142 143 145 151 154 172 177 180 190 195 204

[30] 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.

[31] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 631-644, Feb. 2015.

[32] W. Sun and Q. Du, "Hyperspectral band selection: a review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118-139, Jun. 2019.

[33] W. Sun, J. Peng, G. Yang, and Q. Du, "Fast and latent low-rank subspace clustering for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3906-3915, Jun. 2020.

[34] G. Zhu, Y. Huang, J. Lei, et al., "Unsupervised hyperspectral band selection by dominant set extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 227-239, Jan. 2016.

[35] J. Tschanneral, J. Ren, P. Yuen, et al., "MIMR-DGSA: Unsupervised hyperspectral band selection based on information theory and a modified discrete gravitational search algorithm," *Inform Fusion.*, vol. 51, pp. 189-200, Jan. 2019.

[36] Y. Yuan, X. Zhang, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 51-64, Jan. 2017.

[37] C.-I. Chang, S. Wang, K.-H. Liu, M.-L. Chang, and C. Lin, "Progressive band dimensionality expansion and reduction via band prioritization for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 3, pp. 591-614, Sep. 2010.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[39] Y. Zhan, D. Hu, H. Xing, and X. Yu, "Hyperspectral band selection based on deep convolutional neural network and distance density," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2365-2369, Dec. 2017.

[40] 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.

[41] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognit.*, vol 63, pp. 371-383, Mar. 2017

[42] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, "Gated and axis-concentrated localization network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 179-192, Jan. 2020.

[43] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985-2000, Mar. 2020.

[44] J. Tschannerl, et al., "Segmented autoencoders for unsupervised embedded hyperspectral band selection," *2018 7th European Workshop on Visual Information Processing (EUVIP)*, Tampere, 2018, pp. 1-6.

[45] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks,"*Science*, vol. 344, no. 6191, pp. 1492-1496, Jun. 2014.

[46] X. Lu, W. Zhang, and J. Huang, "Exploiting embedding manifold of autoencoders for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1527-1537, Mar. 2020.

[47] C. Zhao, X. Li, and H. Zhu, "Hyperspectral anomaly detection based on stacked denoising autoencoders," *Proc. SPIE*, vol. 11, no. 4, 2017, Art. no. 042605.

[48] Jang, Eric, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[49] Abid, Abubakar, Muhammad Fatih Balin, and James Zou, "Concrete autoencoders for differentiable feature selection and reconstruction," *arXiv preprint arXiv:1901.09346*, 2019.

[50] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Oct. 2010.

[51] J. Feng, J. Chen, Q. Sun, et al., "Convolutional Neural Network Based on Bandwise-Independent Convolution and Hard Thresholding for Hyperspectral Band Selection," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2020.3000725.

[52] Q. Wang, F. Zhang, and X. Li, "Hyperspectral Band Selection via Optimal Neighborhood Reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol 58, no. 12, pp. 8465-8476, Dec. 2020.