

# Dropout Concrete Autoencoder for Band Selection on Hyperspectral Image Scenes

Lei Xu<sup>1</sup>, Mete Ahishali<sup>1</sup>, and Moncef Gabbouj<sup>1</sup>, *Fellow, IEEE*

**Abstract**—Deep learning-based informative band selection methods on hyperspectral images (HSIs) have recently gained intense attention to eliminate spectral correlation and redundancies. However, existing deep learning-based methods either need additional postprocessing strategies to select the descriptive bands or optimize the model indirectly due to the parameterization inability of discrete variables for the selection procedure. To overcome these limitations, this work proposes a novel end-to-end network for informative band selection. The proposed network, named Dropout concrete autoencoder (CAE), is inspired by advances in the CAE and Dropout feature ranking (Dropout FR) strategy. Unlike traditional deep learning-based methods; the Dropout CAE is trained directly given the required band subset, eliminating the need for further postprocessing. The experimental results in four HSI scenes show that the Dropout CAE achieves substantial and effective performance levels that outperform competing methods. The code is available at <https://github.com/LeiXuAI/Hyperspectral>

**Index Terms**—Autoencoders, deep-learning.

## I. INTRODUCTION

**H**YPERSPECTRAL images (HSIs) captured by hyperspectral remote sensing imaging spectrometers [1] cover a wide and continuous range of the electromagnetic spectrum with multiple spectral bands. Due to this characteristic, HSIs contain enormous information utilizing its various applications, such as in precision agriculture [2], mineral detection [3], and landscape classification [4]. However, massive spectral bands of HSIs imply information redundancy, which leads to the “Hughes phenomenon” [5], computational complexity, and higher storage capacity [6]. Considering the difficulty in selecting prominent wavelengths before capture [6], it is indispensable to develop band selection algorithms for subsequent tasks with HSIs.

Various band selection methods have been proposed to deal with the band redundancy problem, such as ranking-based methods [7], clustering-based methods [8], [9], and searching-based methods [6]. Ranking-based methods are unsupervised approaches that explore specific criteria to rank frequency bands based on the distinctive information of each band. Clustering-based methods [8], [9], [10] usually try to group

relevant bands in selected subsets by computing similarity matrices. These search-based methods use specific search strategies to find the optimal subset with the most informative bands [10].

Deep learning-based methods have been extensively explored to address the band redundancy problem, as shown in [11], [12], and [13]. In [12], an unsupervised end-to-end network was proposed for band selection with a dual-attention mechanism. Feng et al. [5] proposed another end-to-end unsupervised convolutional neural network that combines band selection, feature extraction, and classification. Moreover, studies in [5], [11], [13], and [14] have proposed various autoencoder (AE) models for the band selection task, where after HSI reconstruction with the selected bands, a classifier, such as support vector machine (SVM) or KNN, is used for classification, and the final evaluation is performed on the classification results. Cai et al. [11] proposed an end-to-end AE-based framework, BS network, based on fully connected networks (BS-Net-FC) and convolutional networks (BS-Net-Conv) with an attention mechanism. The BS network explores a band attention module to explicitly model nonlinear interdependencies between the spectral bands [11] with learned weights. Finally, a reconstruction network restores the original spectral bands with the selected number of reweighted bands. Ahishali et al. [13] proposed another AE-based band selection model, self-representation learning with sparse 1-D-operational autoencoder (SRL-SOA). The SRL-SOA model consists of a single 1-D-operational layer encoder with generative neurons for mapping and a self-representation pixelwise decoder for reconstruction.

Concrete AE (CAE) is an unsupervised embedded feature selection method. The CAE inspired by the concrete distribution [15], [16] aims to learn an informative subset and reconstruct the input data from this subset simultaneously. The concrete distribution is introduced for reparameterizations of discrete random variables as of continuous random variables while optimizing the stochastic computation graph via gradient descent [15]. CAE has been extended for band selections in HSI scenes [17], [18]. For example, Sun et al. [17] proposed a Gumbel-Softmax-based CAE and an information entropy criterion for the selection of the optimal band subset. The Gumbel-Softmax distribution [16] can transform a discrete weight matrix into continuous variables for the optimization of selected subsets during the backpropagation of local optimal solutions. Finally, the information entropy criterion searches for a global optimal band subset. In [18], a novel

Received 10 January 2025; revised 6 March 2025 and 14 April 2025; accepted 21 April 2025. Date of publication 25 April 2025; date of current version 16 May 2025. This work was supported by the NSF-Business Finland through the Project Advanced Machine Learning and AI for Industrial Applications (AMALIA). (*Corresponding author: Lei Xu.*)

The authors are with the Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland (e-mail: lei.xu@tuni.fi).

Digital Object Identifier 10.1109/LGRS.2025.3564478

stochastic gate was proposed as a differential layer in the AE-based architecture for a parameterization process based on a Gaussian-based relaxation of Bernoulli variables. The stochastic gate is learnable for an optimal band subset selection without postprocessing for a global optimal result.

Although the methods mentioned above have achieved impressive performance, common limitations still exist in these works. For instance, the nonlinear relation of bands lacks investigation [13] due to the linear convolutional operations as in [5] and [17]. In addition, the required band subset is not optimized directly by the model but is optimized using an approximation of learned weights ranking [11] or band entropy ranking [12]. Moreover, computational complexity increases dramatically as the input scale is larger [11]. In this work, our motivation comes from the concrete relaxation [15] and the Dropout feature ranking (Dropout FR) strategy [19] for HSI band selection tasks. In summary, the main contributions of this work are as follows.

- 1) Unlike existing methods [11], [12], [13], the proposed method can select informative HSI bands without any postprocessing step.
- 2) The architecture of the proposed method is compact and computationally efficient. This is a pioneer study integrating the Dropout strategy with CAEs in an end-to-end fashion for HSI band selection.
- 3) The proposed approach can exploit existing nonlinear dependencies within spectral bands, thanks to the concrete selector.
- 4) The model is optimized and converges directly to a fixed optimal subset revealing the most descriptive spectral bands. Therefore, the proposed approach is fundamentally different from the competing deep neural network (DNN)-based methods [11], [13], [17].

The remainder of this work is organized as follows. First, Section II provides the presentation of the proposed Dropout CAE. The datasets, the details of the implementation, and the experimental results are given in Section III. Finally, we conclude this work in Section IV.

## II. PROPOSED METHOD

In this section, we first present the objective of HSI band selection tasks, the principles of concrete distribution, the Dropout FR, and the Dropout CAE in detail. Next, the pseudo-code of the proposed method is provided at the end of this section.

### A. Band Selection on HSI Using Dropout CAE

The band selection task aims to abandon redundant bands and select an optimal band subset to represent the original HSI. To achieve the target, the Dropout CAE consists of a concrete selector and a fully connected decoder with two layers. The concrete selector takes a 2-D HSI matrix  $\mathbf{x} \in \mathbb{R}^{N \times d}$  as input, where  $N$  is the number of pixels and  $d$  is the number of spectral bands. Then, it samples  $d$ -dimensional binary concrete random variables  $\mathbf{m} \in \mathbb{R}^{d \times 1}$ ,  $\mathbf{m} \in \{0, 1\}^d$  using the Gumbel-Softmax distribution [15] and the Dropout FR strategy [19], [20]. The output of the concrete selector is calculated by  $\mathbf{x} \odot \mathbf{m}$ , which merely retains  $k \ll d$  number of spectral bands

as input to the decoder. The output of the decoder  $f_\theta(\cdot)$  is  $\hat{\mathbf{x}} \in \mathbb{R}^{N \times d}$ , reconstructed HSI matrix.

The Dropout CAE loss function is defined as

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \log p(\mathbf{x}_i | f_\theta(\mathbf{x}_i \odot \mathbf{m}_i)) + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^d m_{ij} \quad (1)$$

where  $\mathbf{m}_i \sim q_\theta(\mathbf{m})$ ,  $q_\theta(\mathbf{m})$  is a variational mask distribution defined as  $q_\theta(\mathbf{m}) = \prod_{j=1}^d q(m_j | \theta_j) = \prod_{j=1}^d \text{Bern}(m_j | \theta_j)$ , the Dropout rate of feature  $j$  is denoted as  $\theta_j$ , and  $\lambda$  is the regularization hyperparameter.

### B. Concrete Distribution

The concrete random variables are defined as a continuous relaxation of discrete random variables [15], [16], which are introduced to address the parameterization inability of discrete random variables during loss propagation by gradient descent. The construction of concrete random variables is motivated by the Gumbel-Max trick [21] sampling from a discrete distribution with *argmax*. The discrete distribution [15] is represented as one-hot vectors  $d \in \{0, 1\}^n$  and  $\sum_{k=1}^n d_k = 1$ . The Gumbel-Max trick cannot be directly used for gradient descent because the *argmax* is a nondifferentiable operation. The softmax function is then introduced to replace the *argmax* for a continuous relaxation of a one-hot vector.

The Gumbel-Softmax distribution as a novel concrete distribution has a closed-form density on the simplex defined with location parameters  $\alpha \in (0, \infty)^n$  and a temperature parameter  $\tau \in (0, \infty)$  [15].  $X \sim \text{Concrete}(\alpha, \tau)$  depicts that  $X$  has the concrete distribution. Then, each element  $X_k$  is sampled as

$$X_k = \frac{\exp((\log \alpha_k + G_k)/\tau)}{\sum_{i=1}^n \exp((\log \alpha_i + G_i)/\tau)} \quad (2)$$

where  $G_k \sim \text{Gumbel}$  independent identically distributed (i.i.d.). Such computation achieves a random probability vector summing to 1. As the temperature parameter  $\tau \rightarrow 0$ ,  $X_k$  is smoothly annealed to the computation of the discrete *argmax*, which means that the Gumbel-Softmax distribution can obtain near one-hot samples with a proper temperature  $\tau$  setting schedule [16].

The encoder part of the CAE architecture consists of a concrete selector layer based on the concrete distribution. Consequently, the selector layer samples a concrete random variable using a temperature parameter  $\tau \in (0, \infty)$  and parameters  $\alpha_k \in (0, \infty)$  for continuous relaxation of a one-hot vector [15], [20].

### C. Dropout Feature Ranking

Bernoulli distribution is a special case of the Gumbel-Max trick with two states of the discrete random variable on  $\{0, 1\}^2$ . When the Gumbel-Softmax trick is implemented on the Bernoulli distribution for a binary concrete random variable, the sampled element  $X \in (0, 1)$  [15] is depicted as

$$X = \frac{1}{1 + \exp(-(\log \alpha + L)/\tau)} \quad (3)$$

where  $L \sim \text{Logistic}$ . When the temperature parameter  $\tau$  follows a proper schedule approach to 0, the output is  $X = 1$  with the probability of  $\alpha/(1 + \alpha)$ .

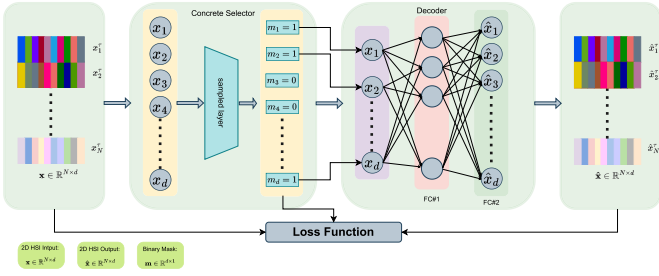


Fig. 1. Architecture of the proposed Dropout CAE.

Variational Dropout as a regularization technique is initially proposed to solve the overfitting problem in deep learning models [22]. The Bernoulli distribution is utilized in the variation Dropout strategy as a Dropout mask in deep learning models. The Dropout mask vector  $\mathbf{m}$  can stochastically determine whether the hidden node in a layer is retained or dropped [19] using a learnable Dropout rate that indicates the importance of the feature, i.e., a lower Dropout rate means a more representative feature.

#### D. Dropout CAE

The general architecture of the Dropout CAE is shown in Fig. 1. The network consists of a concrete selector and a decoder. The concrete selector aims to identify  $k$  number of most informative spectral bands using a learned mask  $\mathbf{m} \in \mathbb{R}^{d \times 1}$ ,  $\mathbf{m} \in \{0, 1\}^2$ . The decoder part comprises two fully connected layers that perform unsupervised reconstruction based on the chosen bands.

The reconstruction performed by the decoder is formulated as

$$\hat{\mathbf{x}} = f_{\theta}(\mathbf{x} \odot \mathbf{m}) \quad (4)$$

where  $\hat{\mathbf{x}} \in \mathbb{R}^{N \times d}$ . The loss function can be further defined as

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^d x_{ij} \log(\hat{x}_{ij}) + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^d m_{ij} \quad (5)$$

where  $\hat{x}_{ij}$  is reconstructed HSI pixel. The first term is cross-entropy loss, and the second one is for regularization. The loss makes the concrete selector learn the importance of band features using Dropout rates and makes the network converge to an optimal band subset (Algorithm 1).

### III. EXPERIMENTAL EVALUATION

In this section, we first describe the datasets used in the experiments and provide implementation details, followed by comparative evaluations.

#### A. Datasets

In this work, we have evaluated the proposed network using four HSI scenes: Indian Pines (IP) [23], PaviaU (PU) [24], Salinas (SA) [25], and KSC [26]. *IP scene* [23] is captured by the AVIRIS sensor with  $145 \times 145$  pixels and 224 spectral bands. The number of bands for the experiment is 200 after removing bands that cover the water absorption regions. The number of pixels for training from this scene

#### Algorithm 1 Pseudo-Code of Dropout CAE

**Input:** 2D HSI matrix  $\mathbf{x} \in \mathbb{R}^{N \times d}$ , the desired band number  $k$ , the temperature parameter  $\tau$ , learnable Dropout rate  $\theta$ , regularization  $\lambda$ , number of epochs  $C$ .

**Output:** Indices of the selected  $k$  number of bands.

- 1 Pre-process the dataset for training;
  - 2 Initialize  $\theta$ ,  $\tau$ ;
  - 3 **for**  $c \in \{1, \dots, C\}$  **do**
  - 4     Get the index of the  $k$  lowest learning rate  $\theta$  using the concrete selector layer for a binary mask  $\mathbf{m}$ ;
  - 5     Reconstruct the input HSI with the selected  $k$  spectral bands as  $\hat{\mathbf{x}}$  using the decoder as Eq. (4);
  - 6     Compute the updated model weights using ADAM optimizer and the loss function Eq. (5);
  - 7     Adjust the temperature parameter  $\tau$  with a schedule;
  - 8 **end**
- Return:** decoder  $f_{\theta}(\cdot)$  and binary concrete parameters

is 10249. *PU scene* [24] is captured by the ROSIS sensor with  $610 \times 610$  pixels and 103 spectral bands. It has a resolution of 1.3 m and nine classes with 42776 training pixels. *SA* [25] is captured by AVIRIS sensors with 512 lines by 217 samples. This scene contains 16 classes, and 204 bands are used for this work. There are 54129 pixels available for training. *KSC scene* [26] (KSC) is captured by the AVIRIS sensor with 224 bands and 5211 pixels for training. The mean imbalance ratios [27] of these scenes are 21.67, 8.35, 5.37, and 3.25 for IP, PU, SA, and KSC, respectively.

The original HSI scenes contain two spatial dimensions and one spectral dimension, denoted as  $\mathbf{x}_{\text{ori}} \in \mathbb{R}^{h \times w \times d}$ . Then, each original HSI scene is transformed into a 2-D HSI matrix  $\mathbf{x} \in \mathbb{R}^{N \times d}$  for training.  $N$  is the number of pixels for training, and  $d$  is the number of spectral bands. The pixel values of  $\mathbf{x} \in \mathbb{R}^{N \times d}$  are normalized to the range [0, 1].

#### B. Implementation

1) *Settings:* The Dropout CAE is implemented with PyTorch [28] on a Nvidia GPU cluster platform. The number of hidden neurons in the decoder part is 128. The hyperparameters for training the model are set as follows: the optimizer is ADAM, the learning rate is 0.001, and  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We adopt the multistep learning rate strategy in PyTorch [28], by which two milestones, 15 and 30, are set with  $\gamma = 0.1$ . The  $\lambda = 0.005$  in (5). The corresponding hyperparameters of the competing methods are set to the default values.

In the experiments, the same samples annotated for the classification are used to train the Dropout CAE model. Then, the SVM [11], [29] classifier is used for performance evaluation with the selected band subset. We randomly select 10% annotated samples from each data scene to train the SVM classifier and 90% samples to test. We run the classification process ten times on each data scene independently for a more fair and robust comparison.

TABLE I  
OVERALL COMPARISON OF THE COMPETING AND THE DROPOUT CAE BAND SELECTION METHODS

Dataset	Indian Pines (25 bands)				Salinas (20 bands)				PaviaU (15 bands)				KSC (15 bands)			
	OA	AA	Kappa	avgEn	OA	AA	Kappa	avgEn	OA	AA	Kappa	avgEn	OA	AA	Kappa	avgEn
SRL-SOA Q=1	0.8070	0.7662	0.7793	<b>4.8855</b>	0.9280	<b>0.9640</b>	0.9197	4.4130	0.9106	0.8838	0.8807	4.2748	0.8664	0.8032	0.8511	4.3004
SRL-SOA Q=3	0.7779	0.7433	0.7448	4.8817	0.9281	0.9627	<b>0.9198</b>	4.5929	<b>0.9190</b>	0.8912	0.8920	4.3106	0.8674	0.8059	0.8521	4.2147
SRL-SOA Q=5	0.7972	0.7617	0.7678	<b>4.8948</b>	0.9258	0.9602	0.9173	<b>4.7771</b>	<b>0.9181</b>	0.8897	0.8908	4.3972	<b>0.8878</b>	0.8308	<b>0.8750</b>	4.3760
SpaBS	0.6298	0.5400	0.5688	4.6911	0.9032	0.9359	0.8919	4.6811	0.8526	0.8126	0.7998	4.1821	0.8391	0.7642	0.8205	<b>4.5898</b>
EGCSR	<b>0.8072</b>	0.7737	<b>0.7795</b>	4.7736	0.9215	0.9592	0.9125	4.5756	0.8624	0.8345	0.8143	4.3927	0.8803	0.8175	0.8666	4.2133
ISSC	<b>0.8123</b>	0.7643	<b>0.7853</b>	4.8107	0.9304	<b>0.9661</b>	<b>0.9224</b>	4.6025	0.9149	0.8891	0.8865	4.3532	<b>0.8920</b>	0.8373	<b>0.8797</b>	4.4991
BS-Net-FC	0.6945	0.7439	0.7080	4.6850	0.9551	0.9174	0.9080	4.7332	0.8761	0.9066	0.8760	4.4205	0.8045	<b>0.8765</b>	0.8620	4.0017
BS-Net-Conv	0.5670	0.6427	0.5920	4.6304	0.9559	0.9178	0.9080	4.5674	0.8881	0.9106	0.8810	<b>4.5012</b>	0.7723	0.8473	0.8300	<b>4.5522</b>
Dropout CAE ( $T_1$ )	0.7701	<b>0.7786</b>	0.7480	4.8532	<b>0.9611</b>	0.9252	0.9166	4.4185	0.8949	<b>0.9215</b>	<b>0.8955</b>	4.4260	0.8003	0.8701	0.8553	4.4188
Dropout CAE ( $T_2$ )	0.7600	<b>0.7827</b>	0.7527	4.7781	<b>0.9614</b>	0.9271	0.9187	<b>4.7421</b>	0.8991	<b>0.9226</b>	<b>0.8970</b>	<b>4.4378</b>	0.8100	<b>0.8753</b>	0.8610	4.1324
Dropout CAE ( $T_3$ )	0.7146	0.7629	0.7300	4.7346	0.9570	0.9170	0.9074	4.5262	0.8865	0.9115	0.8820	4.4155	0.7628	0.8381	0.8196	4.0208
All Bands	0.7965	0.7244	0.7670	4.7480	0.9342	0.9663	0.9266	4.4791	0.9438	0.9234	0.9252	4.3590	0.9127	0.8677	0.9027	4.1031

2) *Evaluation Metrics*: We adopt three quantitative metrics: overall accuracy (OA), average accuracy (AA), and the kappa coefficient (Kappa) [11], [13] for the classification performance of the reconstructed HSI. In addition, the average entropy (avgEn) is calculated based on the avgEn of all bands or chosen bands, which is used to evaluate the effectiveness of the chosen bands. Except for Dropout CAE, BS-Net-FC, and BS-Net-Conv, the final evaluation results from other methods are averaged from the ten runs.

3) *Annealing Schedule*: The Dropout CAE model is optimized with respect to the Dropout rate  $\theta$ , which is highly affected by the setting of the temperature parameter  $\tau$ . Regardless of whether  $\tau$  is high or low, the concrete selector layer converges to a suboptimal informative band subset with a fixed temperature. Therefore, the temperature parameter  $\tau$  is set as a schedule that can gradually decay at each epoch according to a first-order exponential decay as [20]

$$\tau = \tau_0(\tau_C/\tau_0)^{B/(N \times C)} \quad (6)$$

where  $\tau_0$  is the start temperature holding a higher value,  $\tau_C$  is the final temperature with a lower value,  $C$  is the total number of epochs,  $B$  is the batch size, and  $N$  is the total number of pixels for each scene. The annealing schedule begins with the  $\tau_0$  and smoothly decays to  $\tau_C$ . In this work, we set three different annealing schedules for parameters as an ablation study to investigate the effect on the results. These settings are indicated as  $T_1$ ,  $T_2$ , and  $T_3$ , where  $T_1$ :  $\tau_0 = 1$ ,  $\tau_C = 0.001$ ,  $C = 40$ , and  $B = 1$ ;  $T_2$ :  $\tau_0 = 1$ ,  $\tau_C = 0.001$ ,  $C = 200$ , and  $B = 256$ ; and  $T_3$ :  $\tau_0 = 1$ ,  $\tau_C = 0.01$ ,  $C = 200$ , and  $B = 32$ .

### C. Comparisons

We have used five different competing band selection methods for performance comparisons, including SRL-SOA [13] with three  $Q$  values (polynomial approximation order), SpaBS [30], EGCSR [31], ISSC [8], and BS network (BS-Net-FC and BS-Net-Conv) [11]. In addition, three types of annealing schedules are adapted for training under the same conditions to validate the effect of the temperature parameter  $\tau$  with the Dropout CAE model. We selected 25 bands in the IP scene, 20 bands in the SA scene, 15 bands in the PU scene, and 15 bands in the KSC scene with all methods. The top two results are highlighted in bold on each data scene. The overall quantitative results of the four data scenes are listed

TABLE II  
COMPUTATIONAL COMPLEXITY OF METHODS OVER IP SCENE

Model	FLOPs	Params	Run-time (ms)
SRL-SOA (Q=1)	612	800	2.27e-03
SRL-SOA (Q=3)	1.84K	2400	2.29e-03
SRL-SOA (Q=5)	3.06K	4000	2.38e-03
BS-Net-FC	8.25G	154.42K	0.63e-03
BS-Net-Conv	3756.71G	591.58K	1.68e-03
Dropout CAE	0.61G	29.33K	0.62e-03

in Table I. Moreover, the complexity comparison of the deep learning-based methods in the IP dataset is demonstrated in Table II.

As shown in Table I, the Dropout CAE has achieved the best and second-best AA on the IP scene with configurations  $T_2$  and  $T_1$  of 0.7827 and 0.7786, respectively. Similarly, for the SA scene, the Dropout CAE obtained the best OA of 0.9614 and the second-best OA of 0.9611. Compared to other methods on the PU, our proposed model has received the best AA and the second-best AA with  $T_2$  and  $T_1$  as 0.9926 and 0.9215 separately. The best Kappa is also achieved by the Dropout CAE ( $T_2$ ). For the KSC scene, the Dropout CAE ( $T_2$ ) has the best AA. The avgEn using Dropout CAE ( $T_2$ ) is higher than that of all bands in four HSI scenes. In addition, it is observed that our proposed method can achieve better AA metric performance than OA on highly imbalanced scenes, e.g., IP and PU, compared to competing methods. Table II demonstrates the computational complexity with FLOPs, the number of parameters, and run time (ms/sample). As shown, the computational complexity and the number of parameters of the proposed method are lower than those of BS-Net-FC and BS-Net-Conv. Additionally, the run time of SpaBS, EGCSR, and ISSC are 49.47,  $8.51e^{-02}$ , and  $5.32e^{-02}$  ms, respectively. Hence, the run time of the proposed method is the lowest.

In summary, the proposed Dropout CAE can select informative bands effectively without any postprocessing step. Unlike other state-of-the-art (SOTA) works, the proposed method can converge to a fixed optimal subset in the same training environment, which is beneficial for downstream tasks. The performance of downstream classification tasks with the selected band subsets can compete with other SOTA works under quantitative evaluation. In addition, the computational efficiency of the proposed method is outstanding compared

to other deep learning-based methods. The reason is that the computational complexity is highly dependent on the number of required bands. Hence, this method can be a good choice for large-scale datasets with downstream tasks.

#### IV. CONCLUSION

In this work, we propose a novel method named Dropout CAE to reparameterize the discrete random variables for HSI band selection. We first utilize the variational Dropout strategy to exploit the importance of each frequency band for HSI scenes. To bridge the gap between the discrete band information and the reparameterization of the discrete random variables, we introduce the variational Dropout strategy in binary concrete distribution enabling the Dropout CAE model to optimize the model weights and learnable Dropout rates directly. An extensive set of experiments on four different scenes shows that the proposed method outperforms the competing methods in the HSI band selection task. Future works based on the Dropout CAE will focus on an in-depth study for an extensive set of evaluations comparing to high-performance traditional algorithms with high interpretability.

#### REFERENCES

- [1] M. O. Ngadi and L. Liu, *Hyperspectral Image Processing Techniques*. Amsterdam, The Netherlands: Elsevier, Dec. 2010.
- [2] S. G. Bajwa, P. Bajcsy, P. Groves, and L. F. Tian, "Hyperspectral image data mining for band selection in agricultural applications," *Trans. ASAE*, vol. 47, no. 3, pp. 895–907, 2004.
- [3] K. Siebels, K. Goita, and M. Germain, "Estimation of mineral abundance from hyperspectral data using a new supervised neighbor-band ratio unmixing approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 6754–6766, Oct. 2020.
- [4] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.
- [5] J. Feng et al., "Convolutional neural network based on bandwise-independent convolution and hard thresholding for hyperspectral band selection," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4414–4428, Sep. 2021.
- [6] G. Morales, J. Sheppard, R. Logan, and J. Shaw, "Hyperspectral band selection for multispectral image classification with convolutional networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
- [7] 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, Jun. 1999.
- [8] W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2784–2797, Jun. 2015.
- [9] 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.
- [10] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5028–5039, Jun. 2021.
- [11] 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.
- [12] S. K. Roy, S. Das, T. Song, and B. Chanda, "DARecNet-BS: Unsupervised dual-attention reconstruction network for hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2152–2156, Dec. 2021.
- [13] M. Ahishali, S. Kiranyaz, I. Ahmad, and M. Gabbouj, "SRL-SOA: Self-representation learning with sparse 1D-operational autoencoder for hyperspectral image band selection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2296–2300.
- [14] Y. Liu, X. Li, Z. Hua, C. Xia, and L. Zhao, "A band selection method with masked convolutional autoencoder for hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [15] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–20.
- [16] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2017, pp. 1–13.
- [17] H. Sun, J. Ren, H. Zhao, P. Yuen, and J. Tschannerl, "Novel gumbel-softmax trick enabled concrete autoencoder with entropy constraints for unsupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506413.
- [18] H. Sun, L. Zhang, L. Wang, and H. Huang, "Stochastic gate-based autoencoder for unsupervised hyperspectral band selection," *Pattern Recognit.*, vol. 132, Dec. 2022, Art. no. 108969, doi: [10.1016/j.patcog.2022.108969](https://doi.org/10.1016/j.patcog.2022.108969).
- [19] C.-H. Chang, L. Rampasek, and A. Goldenberg, "Dropout feature ranking for deep learning models," 2017, *arXiv:1712.08645*.
- [20] M. F. Balin, A. Abid, and J. Zou, "Concrete autoencoders: Differentiable feature selection and reconstruction," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 444–453.
- [21] C. J. Maddison, D. Tarlow, and T. Minka, "A sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, 2014, pp. 3086–3094.
- [22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [23] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band AVIRIS hyperspectral image data set: Jun. 12, 1992 Indian pine test site 3," Purdue Univ. Res. Repository, Tech. Rep., 1992, doi: [10.4231/R7RX991C](https://doi.org/10.4231/R7RX991C).
- [24] P. Gamba, "Pavia centre and university," Telecommun. Remote Sens. Lab., Pavia Univ., Italy, Tech. Rep.
- [25] EUDAT. (2014). *EUDAT B2SHARE*. [Online]. Available: <http://b2share.eudat.eu>
- [26] D. Bert and H. Ross. (2016). *AmeriFlux AmeriFlux U.S.-KS2 Kennedy Space Center (SCRUB OAK)*. [Online]. Available: [https://www.ehu.edu/ccwintco/index.php/Hyperspectral\\_Remote\\_Sensing\\_Scenes](https://www.ehu.edu/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes)
- [27] L. Xu, J. Raitoharju, A. Iosifidis, and M. Gabbouj, "Saliency-based multilabel linear discriminant analysis," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10200–10213, Oct. 2022.
- [28] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 1–13.
- [29] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
- [30] S. Li and H. Qi, "Sparse representation based band selection for hyperspectral images," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2693–2696.
- [31] Y. Cai, Z. Zhang, Z. Cai, X. Liu, X. Jiang, and Q. Yan, "Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4191–4202, May 2021.