A 3D-CNN Framework for Hyperspectral Unmixing with Spectral Variability

Min Zhao, Shuaikai Shi, Jie Chen, Nicolas Dobigeon

This article has been accepted for publication in a future issue of this journal, but has not been fully edited.
Content may change prior to final publication.
Citation information: DOI 10.1109/TGRS.2022.3141387, IEEE Transactions on Geoscience and Remote Sensing

A 3D-CNN Framework for Hyperspectral Unmixing with Spectral Variability

Min Zhao, *Student Member, IEEE,* Shuaikai Shi, *Student Member, IEEE,*
Jie Chen, *Senior Member, IEEE,* and Nicolas Dobigeon, *Senior Member, IEEE*

***Abstract*—Hyperspectral unmixing plays an important role in hyperspectral image processing and analysis. It aims to decompose mixed pixels into pure spectral signatures and their associated abundances. The hyperspectral image contains spatial information in neighborhood regions, and spectral signatures existing in the region also have high correlation. However, most autoencoder (AE) based unmixing methods are pixel-to-pixel methods and ignore these priors. It is helpful to add spectral-spatial information into unmixing methods. A recent trend to deal with this problem is to use convolutional neural networks (CNNs). Our proposed framework uses 3D-CNN based networks to jointly learn spectral-spatial priors. Moreover, previous AE-based unmixing methods use fixed spectral signatures for each pure material. In our work, we use a carefully designed decoder to cope with the endmember variability issue, and variational inference strategy is applied to add uncertainty property into endmembers. To avoid over-fitting, we use structured sparsity regularizers to the encoder networks, and $\ell_{2,1}$-loss is added to the estimated abundances to guarantee the sparseness. Experimental results on both simulated and real data demonstrate the effectiveness of our proposed method.**

***Index Terms*—Hyperspectral imaging, unmixing, endmember variability, 3D-CNN, structured sparsity, weight uncertainty.**

I. Introduction

HYPERSPECTRAL imaging is a rapidly developing field in remote sensing, it incorporates both imaging and spectral techniques [1], [2]. Benefiting from the high spectral resolution, it enables to analyze the materials in a more accurate manner and contribute to a wide range of applications, such as remote surveillance, environment monitoring, and food safety [3], [4]. However, because of the limited spatial resolution of hyperspectral sensors, a pixel of hyperspectral image may capture a scene containing more than one dis-tinct material, which makes a performance degradation for the further analysis of the hyperspectral image. Therefore, hyperspectral unmixing has become a hot topic to deal with the mixed pixel problem. It aims to parse the mixed pixel into a set of pure spectral signatures (named endmembers) and their related fractional percentages (named abundances) [5], [6].

The most common approaches to address hyperspectral unmixing can be divided into three categories, namely, supervised, semi-supervised and unsupervised methods. Supervised methods solve the unmixing task with endmembers known as a prior or extracted from endmember extraction methods. Then the problem builds down to the estimation of abundances [7]. Semi-supervised methods estimate abundances along with selecting proper endmembers from a given spectral library, e.g., by imposing a sparse representation of the measured pixel spectra over the library atoms [8]. Unsupervised approaches, also termed as blind unmixing methods, simultaneously estimate the endmembers and corresponding abundances from the hyperspectral image with the number of endmembers known as a prior [9]. The method proposed in this work belongs to this latter category.

A. Motivation

Thanks to the spatial correlation and spectral similarity in the neighboring pixels, properly using potential spectral-spatial priors can effectively improve the unmixing performance. Conventional unmixing algorithms usually exploit these priors by plugging smoothness regularizers into the optimization problem [7], [10]. Conversely, most deep learning (DL) based blind unmixing frameworks are pixel-wise models and do not fully exploit the spatial and spectral structures of hyperspectral images. Recently, convolutional neural networks (CNNs) have shown promising performance in image processing tasks, such as image classification [11], [12] and object detection [13], [14]. More recently, 3D-CNNs have shown to be a relevant tool to extract meaningful information in both spatial and spectral domains, with promising results with respect to a classification task [15], [16]. In this work, we propose to leverage on the 3D-CNN based framework to fully exploit spectral-spatial features inherent of hyperspectral images to conduct unmixing.

In addition, selecting a proper mixing model is also a non-trivial task. Many mixing models have been proposed in the past decade to explain the relationship between endmembers and abundance fractions [17], [18]. The linear mixing model (LMM) is the most widely used model thanks to its computational simplicity and its ease of physical interpretability. However, due to varying illumination conditions, intrinsic variability and multiple scatters among pure materials, the LMM is not valid to accurately describe some real scenes.

To cope with this issue, several nonlinear mixing models have been proposed, including bilinear models [19], the Hapke model [20] and kernel-based models [10], [21]. Nevertheless, these models assume that each endmember can be characterized by a unique spectral signature, thus neglecting any spectral variability affecting the endmembers [18].

B. Our contributions

This paper introduces a CNN-based autoencoder (AE) network to solve the unmixing problem. More precisely, the encoder proposes a 3D-CNN architecture to jointly learn spectral-spatial features of the hyperspectral images. The decoder leverages on the perturbed linear mixing model (PLMM) to explicitly account for endmember variabilities. The main contributions of this paper are summarized as follows:

- A novel 3D-CNN architecture is proposed to conduct spectral unmixing. It can effectively extract spectral-spatial features characterizing the image, which enhances the unmixing performance.
- A decoder is specifically designed to account for endmember variability. Variational Bayesian learning is conducted to derive the probability distribution of the endmembers.
- Structured sparsity is included into the loss function to regularize the structured shape of our encoder networks. In particular, group lasso is conducted on different patterns, namely, filter-wise group, channel-wise group and layer-wise group. This regularizer can avoid over-fitting. A sparsity-promoting regularization is also considered to promote sparse abundances.

The paper is organized as follows. Section II briefly reviews related works dealing with endmember variability and DL-based unmixing approaches. Section III describes the proposed 3D-CNN based framework and the proposed strategy to account for the endmember variability within the proposed architecture. Section V and VI compare the performance of the proposed method with those obtained by state-of-the-art unmixing methods. Finally, conclusions of our work are presented in Section VII.

II. Related Works

A. AE-based hyperspectral unmixing

Nowadays, as an advanced machine learning method, DL has demonstrated promising performance while addressing hyperspectral unmixing. DL-based unmixing approaches can be divided into two main classes, namely, classifier-based models and AE-based models. Classifier-based unmixing methods are supervised approaches, which train the architecture to directly map input spectra to abundance fractions. One limitation of such approaches is that they are supervised and thus need training data with ground-truth, which heavily limits their applicability because of the lack of labeled datasets. Conversely, AE-based unmixing approaches are unsupervised and are not limited the prior availability of training dataset. Some recently proposed AE-based unmixing methods are briefly reviewed in what follows.

Recently, there has been growing interest in proposing AE-like deep architectures specifically dedicated to the unmixing problem. By leveraging on a generative description of the observed hyperspectral images, AE-based unmixing methods allow the end-users to extract the endmember signatures and to estimate the corresponding abundances simultaneously. For example, the strategy in [22] used a regular multiple hidden layer AE framework to conduct hyperspectral unmixing. However, hyperspectral images are usually corrupted by noise or outliers, which may lead to significant performance degradations. To cope with this issue, denoising-oriented unmixing architectures can be derived, see, e.g., [23] or, more recently, [24] where a denoising constraint was included into the network to achieve noise reduction. In [25] the authors followed a different route by deriving a strategy to relevantly select noiseless samples from the training set.

Since observed pixels are expected to contain a few elementary materials, sparse constraints have been also widely considered to regularize DNN-based unmixing methods. Specifically, the work in [26] imposed an ℓ_1-norm to the latent output to generate sparse unmixing results. In [24], an $\ell_{2,1}$-norm constraint was added to the weights of the last layer of the encoder. Stacked nonnegative sparse AEs were also introduced to observe sparse unmixing results [27]. All of the aforementioned AE-based unmixing schemes were designed following the standard LMM. Although exhibiting overall good performance, this model can be inappropriate for some real scenarios, due to the existence of multiple scatters or the diversity of illumination conditions. Several attempts were made to learn the nonlinearity, for instance by designing a dedicated decoder [28], [29]. These works illustrate that DNN shows superior efficiency to address the nonlinear unmixing problem. However, hyperspectral images are also characterized by spatial correlation among pixels. These correlations are widely neglected by DNN-based architectures which only focus on the spectral information and neglect the neighboring correlation.

To fully exploit the spatial information, AE architectures have been also designed to analyze hyperspectral data. For instance, the authors of [30] used 2D-CNN to extract spatial structures present in the hyperspectral images to be unmixed. However, 2D-CNN methods analyze hyperspectral images across bands independently, thus neglecting the spectral nature of the data. Conversely, 3D-CNN are relevant architectures to extract spatial-spectral features from multiple adjacent bands. Some works exploit this ability of modeling spatial-spectral data for hyperspectral image processing tasks, including super-resolution [31], denoising [32] and inpainting [33]. Fig. 2 illustrates the key difference between the 2D and convolution operations. In [34], a 3D-CNN based framework has shown superior performance to incorporate spectral-spatial priors. However, this latest approach assumed that the endmembers were known and fixed a priori and neglected any spectral variability affecting their signatures. Besides, these architectures are shallow and have limited ability to capture spatial information and to mitigate the effects of the noise. Moreover, these methods use batch-normalization to prevent over-fitting, but it requires expensive computing resources and also significantly

increases the time of gradient calculations.

B. Hyperspectral unmixing with spectral variability

Spectral variability has been considered as one of the most important issues in hyperspectral image processing. Several reasons responsible for spectral variability can be identified. First, the illumination conditions may greatly affect the shape and the intensity of the spectral signatures over a unique image or across several images, e.g., in case of multi-temporal acquisitions. Second, the relief of the imaged surfaces, interactions between neighboring pixels or microscopic scatters between the materials may also induce fluctuations in the spectral signatures. Recently, many researches have been conducted to tackle the spectral variability issue, in particular when performing spectral unmixing. These approaches are briefly discussed below.

There are mainly three different classes of methods to cope with the endmember variability problem. The first class relies on several spectral instances of a given endmember to represent each pure material. These instances can be extracted from the image and subsequently clustered in groups usually referred to bundles [35], [36]. These classes are generally built according to the spectral attributes of the pixel signatures to ensure spectral consistency within each endmember bundle while maximizing the spectral dissimilarity between bundles. Several methods proposed to exploit both spatial and spectral information to extract these bundles [36]–[38]. Then each pixel of the hyperspectral image can be unmixed by looking for the most representative endmember in each bundle, leading to the popular multiple endmember spectral mixture analysis (MESMA) [39]. One major limitation of MESMA lies in its high computational complexity due to the numbers of combinations to be tested to identify the most representative endmember spectra. To lighten this complexity, one strategy consists in exploiting the expected sparsity of the pixel [40], [41].

Other major limitations of the bundle-based methods are that no explicit model is derived to describe the spectral variability and they highly rely on the empirical definition of the bundles, which may be limited to describe complex models. To alleviate, many parametric of statistical mixing models incorporating endmember variability have been introduced. The extended linear mixing model (ELMM) uses a linear combination of spatially correlated endmembers to introduce variability [42]. The generalized extended linear mixing model (GELMM) extends ELMM by defining a band-dependent variation of the endmember signatures [43]. The PLMM considers that each endmember is affected by an addictive perturbation, which can capture the varying spectral-spatial variabilities [44]. Conversely, statistical mixing models were also proposed to derive a stochastic description of the variability. Bayesian approaches were applied to learn the distribution of endmembers, such as normal compositional model (NCM) [45] and beta compositional model (BCM) [46]. In [47], a Gaussian mixture model (GMM) was proposed to represent the endmember variability.

Finally, some DL-based frameworks have been also designed to account for the endmember variability. For example, the endmember-guided unmixing network (EG-net) proposed in [48] introduced an additional network to address the endmember variability problem. However, it was a self-supervised method and needs some training data with associated ground-truth, which limited its applicability to real-world scenario. In [49], a variational AE (VAE) was trained to learn the probability distribution of the endmembers. Then the decoder was used as a black box to generate endmembers, and the abundances were estimated by a classical optimization method. Again, one drawback of this work is that it needs training data to learn the generative model.

III. Problem formulation

LMM has been widely employed in unmixing, while it remains inaccurate when facing with spectral variability. Alternatively, PLMM extends the conventional formulation LMM by explicitly defining a parametric model describing the variability. More precisely, it includes a band- and pixel-wise additive endmember perturbations to account for endmember variabilities of low energy. This model has been widely used in various contexts [44] including online [50] or distributed [51] contexts and can account for deviations from the LMM. It can be written as

$$\boldsymbol{r}_i = \sum_{p=1}^{P} (\boldsymbol{e}_p + \boldsymbol{d}_{p,i}) a_{p,i} + \boldsymbol{n}_i \tag{1}$$

where $\boldsymbol{r}_i \in \mathbb{R}^L$ is the measured spectrum in L spectral bands associated with the ith pixel of the hyperspectral image, $\boldsymbol{e}_p$ denotes the pth nominal endmember signature. This nominal signature is expected to be locally perturbed due to various natural phenomena such as illumination variations and intrinsic variability. These variations affect the nominal endmember signature in each pixel individually and are modeled as an additive perturbation, denoted $\boldsymbol{d}_{p,i}$ for the pth endmember perturbation in the ith pixel. The coefficient $a_{p,i}$ represents the abundance of the ith pixel related to the pth endmember and $\boldsymbol{n}_i$ is a zero-mean Gaussian noise. In the sequel, we assume that the number P of endmembers is known or can be estimated, e.g., using HySime [52]. Using standard matrix notations, the P nominal endmember spectra forms the matrix denoted $\boldsymbol{E} \in \mathbb{R}^{L \times P}$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N] \in \mathbb{R}^{P \times N}$ is the matrix whose ith column $\boldsymbol{a}_i = [a_{1,i}, \ldots, a_{P,i}]^T$ is composed of the abundance coefficients associated with the ith pixel. The image to be unmixed is assumed to be composed of N_r rows and N_c columns resulting in $N = N_r N_c$ pixels whose spectral signatures $\boldsymbol{r}_i$ ($i = 1, \ldots, N$) can be arranged in the 3-dimensional array $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{L \times N_r \times N_c}$ or equivalently, by lexicographically ordering these signatures, in the matrix $\boldsymbol{R} \in \mathbb{R}^{L \times N}$. The set of variability terms $\boldsymbol{d}_{p,i} \in \mathbb{R}^L$ ($p = 1, \ldots, P$, $i = 1, \ldots, N$) are gathered in the 3-dimensional array $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{L \times P \times N}$. Using these notations, PLMM applied for the N pixels can be rewritten in a compact form as

$$\boldsymbol{R} = \boldsymbol{\mathcal{M}} \otimes \boldsymbol{A} + \boldsymbol{N}, \tag{2}$$

where $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{L \times P \times N}$ is a 3-dimensional array of the pixel-wise perturbed endmember signatures, i.e.,

$$\boldsymbol{\mathcal{M}}_{:,:,i} = \boldsymbol{E} + \boldsymbol{\mathcal{D}}_{:,:,i} \triangleq \boldsymbol{M}_i, \ i = 1, \ldots, N \tag{3}$$

and $\otimes$ stands for a specific operator defined by

$$\boldsymbol{\mathcal{M}} \otimes \boldsymbol{A} = [\boldsymbol{M}_1\boldsymbol{a}_1, \ldots, \boldsymbol{M}_N\boldsymbol{a}_N]. \tag{4}$$

The noise matrix $\boldsymbol{N} = [\boldsymbol{n}_1, \ldots, \boldsymbol{n}_N]$ satisfies $\boldsymbol{n}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ $(n = 1, \ldots, N)$ with $\boldsymbol{\Sigma} = \text{diag}(\delta_1^2, \delta_2^2, \cdots, \delta_L^2)$, i.e., the noise is assumed to be pixel-wise independent but with a band-wise signal-to-noise ratio (SNR).

The abundance maps are considered to satisfy two constraints, namely, abundance nonnegative constraint (ANC) and abundance sum-to-one constraint (ASC), which are consistent with the physical interpretation of the abundances. The endmember matrix are also required to satisfy a nonnegativity constraint (ENC). This set of constraints can be expressed as

$$\begin{aligned} &\boldsymbol{A} \succeq \boldsymbol{0}, \quad \boldsymbol{1}_P\boldsymbol{A} = \boldsymbol{1}_N, \\ &\boldsymbol{E} \succeq \boldsymbol{0}, \quad \boldsymbol{\mathcal{M}} \succeq \boldsymbol{0}, \end{aligned} \tag{5}$$

where $\succeq$ denotes an element-wise operator, $\boldsymbol{1}_P \in \mathbb{R}^{1\times P}$ and $\boldsymbol{1}_N \in \mathbb{R}^{1\times N}$ represent all-one row vectors.

IV. Proposed Method

This section details the proposed unmixing method, whose architecture is depicted in Fig. 1. As previously, this method is based on an AE framework able to deal with unsupervised unmixing under endmember variability. Similarly to the most conventional AEs, the architecture is divided into two parts, namely, encoder and decoder. The encoder is designed to map the input pixels into a lower dimensional representation, which can be formulated as

$$\boldsymbol{\mathcal{H}} = \mathsf{E}(\boldsymbol{\mathcal{R}}) \tag{6}$$

where $\mathsf{E}(\cdot) : \mathbb{R}^{L\times Nr\times Nc} \to \mathbb{R}^{P\times Nr\times Nc}$ is a nonlinear function and $\boldsymbol{\mathcal{H}}$ is the compressed image representation that can be reshaped in a matrix form as $\boldsymbol{H} \in \mathbb{R}^{P\times N}$. To take advantage of the spectral-spatial nature of the input data, this part is specifically designed with 3D-CNNs. More details on this encoding step are given in Section IV-A.

The decoder is applied to the lower representation to reconstruct the input data, i.e.,

$$\hat{\boldsymbol{\mathcal{R}}} = \mathsf{D}(\boldsymbol{H}), \tag{7}$$

where $\mathsf{D}(\cdot) : \mathbb{R}^{P\times N} \to \mathbb{R}^{L\times Nr\times Nc}$ is the decoding function. In particular, this function is designed according to the following reconstruction model

$$\hat{\boldsymbol{R}} = \boldsymbol{\mathcal{W}} \otimes \boldsymbol{H}, \tag{8}$$

where $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{L\times P\times N}$ are the decoder weights whose specific form will be defined later in (13), and $\hat{\boldsymbol{R}} \in \mathbb{R}^{L\times N}$ is the reconstructed data. To account for endmember variability, these weights are learnt within a probabilistic framework, as detailed in Section IV-B.

The parameters of the network are estimated by back propagation and using a stochastic gradient descent algorithm. When comparing (8) with the matrix form of PLMM (2), it clearly appears that the decoder mimics the output in agreement with this model. Therefore, after the entire network has been trained, the output of the encoder (i.e., the input of the decoder) can be considered as abundances, the weights of decoder are associated with endmembers. Interested readers are invited to consult [53] for a thorough interpretation of the relation between the decoder structure and mixture models. Thus, a one-to-one correspondence can be drawn between the endmember and abundance matrices on one side, and the compressed representation and decoder weights on the other side. Finally, the target quantities will be estimated as

$$\hat{\boldsymbol{\mathcal{M}}} \leftarrow \boldsymbol{\mathcal{W}} \quad \text{(endmember estimation)} \tag{9}$$

$$\hat{\boldsymbol{A}} \leftarrow \boldsymbol{H} \quad \text{(abundance estimation)}. \tag{10}$$

A. Encoder

The proposed encoder is designed using CNNs with 3D convolution kernels because of their ability to act in both spatial and spectral dimensions. Fig. 2 illustrates the 3D convolution operation and the architecture of the proposed encoder is shown in Fig. 1 (left). It is composed of six CNN layers where the number of filters decreases with the depth of the layer. The first five layers use 3D convolution kernels with spatial size of 3×3 to learn the spatial information. The last layer uses 3D convolution kernel with spatial size of 1×1 to focus on compressing the spectral information. A downsampling of factor 2 is applied to the feature maps in the spectral dimension. Furthermore, instead of using dropout to avoid over-fitting, we rather resort to a structured sparse regularizer, which can not only avoid over-fitting but also reduce the computing burden of the network. More details are presented in Section IV-C. The spectral dimension of the encoder output is set to the number P of endmembers. Leaky ReLU is used as activation function of each layer. Softmax is particularly selected as the activation function of the last layer to ensure that the output satisfies ANC and ASC.

B. Decoder

Archetypical AE-based unmixing architectures usually assume a fixed endmember signature for each pixel and are not able to account for endmember variability. Conversely, in this work, we aim at designing a decoder able to tackle this challenge. As suggested by (9), the decoder weights $\boldsymbol{\mathcal{W}}$ to be learnt stand for the perturbed endmember signatures. To model the uncertainty resulting for the variability, these weights are assumed to be random variables whose posterior distribution writes

$$P(\boldsymbol{\mathcal{W}} \mid \boldsymbol{H}, \boldsymbol{R}) = \frac{P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}})P(\boldsymbol{\mathcal{W}})}{P(\boldsymbol{H}, \boldsymbol{R})}. \tag{11}$$

Unfortunately, deriving this posterior remains intractable. Thus a variational approximation is proposed, which consists in minimizing the Kullback-Leibler (KL) divergence between the posterior in (11) and an instrumental distribution $q(\boldsymbol{\mathcal{W}} \mid \boldsymbol{Z})$ parameterized by $\boldsymbol{Z}$. This can be formulated as the optimization problem

$$\min_{\boldsymbol{Z}} \ \text{KL}\left[q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z}) \| P(\boldsymbol{\mathcal{W}}|\boldsymbol{H}, \boldsymbol{R})\right]. \tag{12}$$

Directly minimizing (12) is still non-trivial due to the stochastic nature of the weights. To alleviate, one resorts to

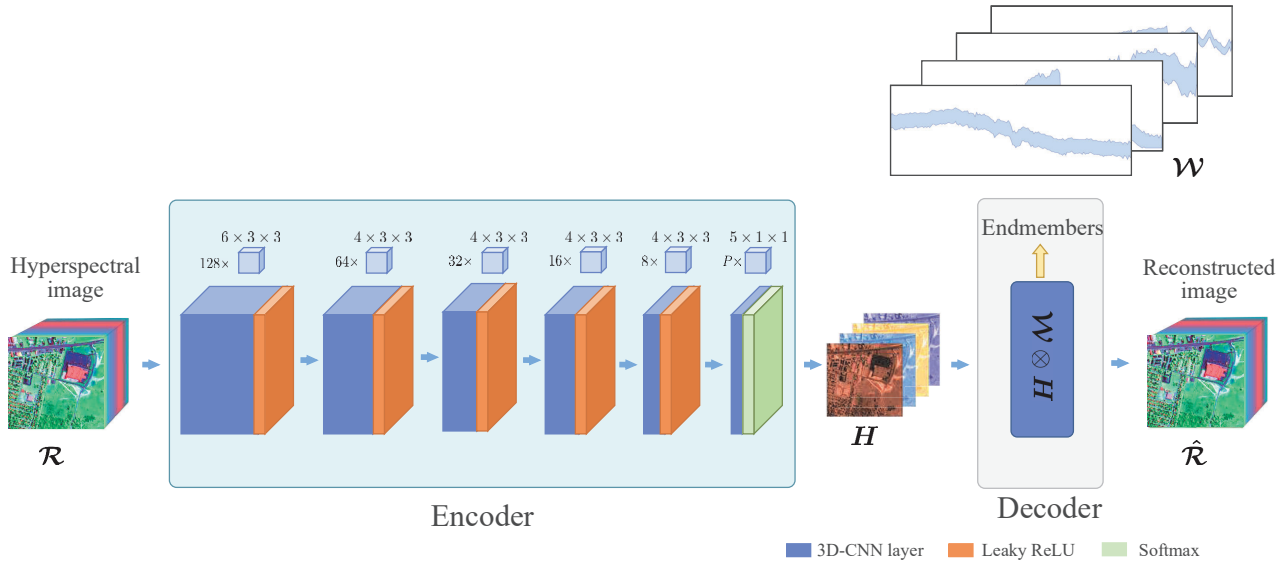

Fig. 1. The architecture of our proposed framework. It consists of two parts, encoder and decoder. The estimated abundances $\boldsymbol{H}$ are the output of encoder, and the extracted endmembers $\boldsymbol{\mathcal{W}}$ are the weights of decoder, which also devotes to the endmember variability task.

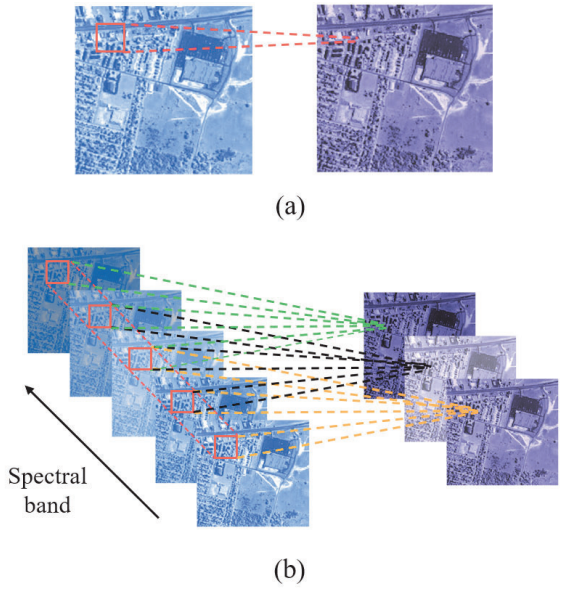

Fig. 2. Schematic illustration of 2D (a) and 3D (b) convolution operations.

the reparameterization trick by defining the mappings ($p = 1, \ldots, P$; $i = 1, \ldots, N$)

$$\mathbf{w}_{p,i} = \boldsymbol{u}_p + \boldsymbol{v}_{p,i}, \tag{13}$$

where $\boldsymbol{u}_p$ and $\boldsymbol{v}_{p,i}$ are considered as random variable with $\boldsymbol{v}_{p,i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{p,i})$. In other words, the variational posterior is assumed to obey a Gaussian distribution, and by assuming independence on pixels and endmembers, can be factorized as

$$q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{p=1}^{P} \mathcal{N}\left(\mathbf{w}_{p,i}; \boldsymbol{u}_p, \boldsymbol{\Sigma}_{p,i}\right) \tag{14}$$

where, with a slight abuse of notations, $\mathcal{N}(\cdot; \boldsymbol{a}, \boldsymbol{B})$ denotes the probability density function of the Gaussian distribution of mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{B}$.

Following the estimation strategy (9), by comparing (13) with the PLMM formulation (1) or, more precisely, the definition of the perturbed endmember signatures (3), the pixel-independent mean vector $\boldsymbol{u}_p$ ($p = 1, \ldots, P$) can be interpreted as the pth endmember nominal signature while the stochastic pixel-dependent component $\boldsymbol{v}_{p,i}$ is expected to capture the endmember variability over pixels. Two strategies are considered to model $\boldsymbol{v}_{p,i}$:

- First, to save computational resources and time, we consider $\boldsymbol{v}_{p,i}$ is band independent, i.e., it can be written as $\boldsymbol{v}_{p,i} = \varepsilon_{p,i}\boldsymbol{\sigma}_{p,i}$, where $\varepsilon_{p,i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In other words this model, referred as 3DCNN-var/I in the sequel, assumed a diagonal covariance matrix $\boldsymbol{\Sigma}_{p,i} = \sigma_{p,i}^2\boldsymbol{I}$ with $\sigma_{p,i}^2 = \log(1 + \exp(\phi_{p,i}))$. The variational parameters write $\boldsymbol{Z} = \{\boldsymbol{U}, \boldsymbol{\Phi}\}$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_P] \in \mathbb{R}^{L\times P}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{P\times N}$ with $[\boldsymbol{\Phi}]_{p,i} = \phi_{p,i}$. These parameters can be updated according to the rules

$$\begin{aligned} \boldsymbol{U} &\leftarrow \boldsymbol{U} - \upsilon\nabla_{\boldsymbol{U}} \\ \boldsymbol{\Phi} &\leftarrow \boldsymbol{\Phi} - \upsilon\nabla_{\boldsymbol{\Phi}} \end{aligned} \tag{15}$$

  where υ is the learning rate.
- Second, to account for correlation across bands, we consider the richer model denoted 3DCNN-var/D defining the perturbation as $\boldsymbol{v}_{p,i} = \mathbf{L}_{p,i}\varepsilon_{p,i}$, where $\mathbf{L}_{p,i}$ is a lower or upper triangular matrix with nonzero entries on the diagonal. The off-diagonal elements define the correlations between bands, and $\varepsilon_{p,i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In this case, the covariance matrix can be derived as $\boldsymbol{\Sigma}_{p,i} = \mathbf{L}_{p,i}\mathbf{L}_{p,i}^{\mathrm{T}}$ where $\mathbf{L}_{p,i}$ is the corresponding Cholesky factor. More details can be found in [54], [55]. The variational

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3141387, IEEE Transactions on Geoscience and Remote Sensing

parameters are $\boldsymbol{Z} = \{\boldsymbol{U}, \boldsymbol{L}\}$ and can be updated using the following rules

$$\begin{aligned} \boldsymbol{U} &\leftarrow \boldsymbol{U} - \upsilon \nabla_{\boldsymbol{U}} \\ \boldsymbol{L} &\leftarrow \boldsymbol{L} - \upsilon \nabla_{\boldsymbol{L}}. \end{aligned} \tag{16}$$

The first strategy is simple, but it ignores the correlation between spectral bands. The second one uses a full-covariance matrix to define the correlation between the bands, at cost of higher computational resources. Both models will be compared in the experimental section to evaluate the trade-off between performance and complexity (see Section V).

C. Overall loss function

The overall loss of the proposed framework is composed of four terms designed to reach better unmixing performance. They are detailed below.

Reconstruction loss – This term is classically introduced to ensure consistency between input pixels and their corresponding reconstructed pixels, i.e.,

$$\mathcal{L}_{\text{data}} = \frac{1}{N} \left\| \boldsymbol{R} - \hat{\boldsymbol{R}} \right\|_{\text{F}}^2. \tag{17}$$

KL loss – The problem (12) can be turned into the maximization of the corresponding evidence lower bound (ELBO), leading to the loss function

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z})} \left[\log \frac{q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z})}{P(\boldsymbol{\mathcal{W}}) P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}})} \right]. \tag{18}$$

This loss is used to train the proposed decoder while accounting for the endmember variability. The loss function (18) can be rewritten as

$$\mathcal{L}_{\text{KL}} = \text{KL}\left[q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z}) \| P(\boldsymbol{\mathcal{W}}) \right] - \mathbb{E}_{q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z})} \left[\log P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}}) \right]. \tag{19}$$

The first term of (19) is a KL term. The weights $\boldsymbol{\mathcal{W}}$ are assigned a separable prior distribution

$$P(\boldsymbol{\mathcal{W}}) = \prod_{i=1}^{N} \prod_{p=1}^{P} \mathcal{N}\left(\mathbf{w}_{p,i}; \bar{\boldsymbol{u}}_p, \bar{\boldsymbol{\Sigma}} \right) \tag{20}$$

where the $\bar{\boldsymbol{u}}_p$ and $\bar{\boldsymbol{\Sigma}}$ are the prior means and covariance matrices. In this work, based on the particular interpretation of the weights (see previous section), the matrix $\bar{\boldsymbol{U}} = [\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_P]$ is chosen as crude estimates of the endmembers, e.g., provided by vertex component analysis (VCA) [56], such that physically-based prior information on the endmembers can be incorporated into the estimation procedure. Thus, the first term in the right-hand side of (19) can be rewritten as

$$\begin{aligned} &\text{KL}\left[q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z}) \| P(\boldsymbol{\mathcal{W}}) \right] = \\ &\frac{1}{2} \sum_{p,i} \left[\log \left(\frac{\bar{\boldsymbol{\Sigma}}}{\boldsymbol{\Sigma}_{p,i}} \right) + \| \boldsymbol{u}_p - \bar{\boldsymbol{u}}_p \|_{\bar{\boldsymbol{\Sigma}}^{-1}}^2 + \text{tr}\left(\bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_{p,i} \right) - L \right]. \end{aligned} \tag{21}$$

The structure of the prior covariance matrix $\bar{\boldsymbol{\Sigma}}$ is selected as follows. For 3DCNN-var/I, we set $\bar{\boldsymbol{\Sigma}} = \lambda \mathbf{I}$, where $\lambda \in [0, 1]$ adjusts the level of endmember perturbation. For the richer model 3DCNN-var/D, $\bar{\boldsymbol{\Sigma}}$ is a symmetric positive definite matrix. The elements of $\bar{\boldsymbol{\Sigma}}$ reduce as they are away from the diagonal, as neighboring bands are usually more correlated. We set $\bar{\boldsymbol{\Sigma}}$ as as a Teoplitz matrix following a 1st order autoregressive structure with the ith element of the 1st row denoted λ^i.

The second term of (19) can be approximated by Monte Carlo sampling

$$\mathbb{E}_{q(\boldsymbol{\mathcal{W}}|\boldsymbol{Z})} \left[\log P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}}) \right] \approx \frac{1}{J} \sum_{j=1}^{J} \log P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}}^{(j)}), \tag{22}$$

In this work, the number of samples are chosen as $J = N$. By writing $P(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{\mathcal{W}}) = P(\boldsymbol{R}|\boldsymbol{\mathcal{W}}, \boldsymbol{H}) P(\boldsymbol{H})$ and assuming that

$$P(\boldsymbol{R}|\boldsymbol{\mathcal{W}}, \boldsymbol{H}) = \mathcal{N}(\boldsymbol{R}; \hat{\boldsymbol{R}}, \boldsymbol{\Sigma}_r), \tag{23}$$

where $\hat{\boldsymbol{R}} = \boldsymbol{\mathcal{W}} \otimes \boldsymbol{H}$, it appears that this second term only depends on data fitting terms similar to (17), up to some constants. Since the reconstruction error has been already considered in the whole loss function, the KL loss boils down to its first term (21).

Structured sparsity loss – Motivated by the fact that redundant information is expected to be propagated through the encoder, a structured sparsity is first introduced to regularize the structure of the proposed encoder. This will have the benefit of not only reducing the computational cost but also avoiding over-fitting. Let S denote the number of layers defining the encoder. The weights of the sth layer are gathered in a 5-dimensional array $\boldsymbol{\Lambda}^{(s)} \in \mathbb{R}^{F_s \times C_s \times D_s \times M_s \times K_s}$ where F_s is the number of filters, C_s is the channel size, and D_s, G_s and K_s stand for the depth, width and height of the filter, respectively. Three levels of group-sparsity are imposed and described below.

- *Filter-wise sparsity*: Using consistent notation, ones denotes $\boldsymbol{\Lambda}^{(s)}_{f,:,:,:,:}$ is the fth filter of the sth layer. Filter-wise sparsity is imposed for each layer, according to

$$\mathcal{L}_{\text{sp-filter}} = \sum_{s=1}^{S} \left(\sum_{f=1}^{F_s} \left\| \boldsymbol{\Lambda}^{(s)}_{f,:,:,:,:} \right\|_{\text{F}} \right). \tag{24}$$

- *Channel-wise sparsity*: Similarly, $\boldsymbol{\Lambda}^{(s)}_{:,c,:,:,:}$ stands for the c-th channel of the sth layer and the group sparsity accross channels writes function on the structured sparsity of channels can be defined as:

$$\mathcal{L}_{\text{sp-channel}} = \sum_{s=1}^{S} \left(\sum_{c=1}^{C_s} \left\| \boldsymbol{\Lambda}^{(i)}_{:,c,:,:,:} \right\|_{\text{F}} \right). \tag{25}$$

- *Layer-wise sparsity*: Finally, the depth of the encoder is regularized by considering the group-sparsity promoting term

$$\mathcal{L}_{\text{sp-layer}} = \sum_{s=1}^{S} \left\| \boldsymbol{\Lambda}^{(s)} \right\|_{\text{F}}. \tag{26}$$

By using these structured group-sparsity regularizers, irrelevant filters and/or channels can be implicitly removed. Moreover, the layer-wise sparsity makes weights get small values to avoid over-fitting.

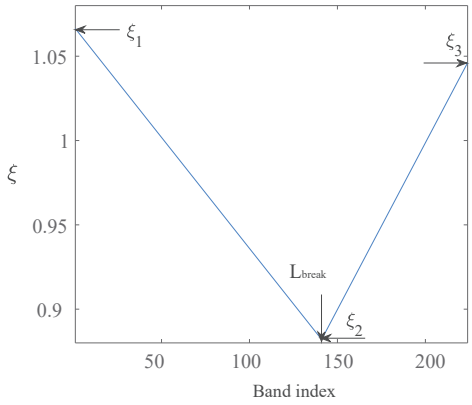

Fig. 3. An example of the randomly generated affine function to mimic endmember variability.

Besides, since only a few endmembers are expected to contributed to each measured pixel spectrum, the abundances are also accompanied by a sparsity promoting regularization [6]. It is formulated as:

$$\mathcal{L}_{\text{sp-abu}} = \sum_{i=1}^{P} \|\boldsymbol{H}_{i,:}\|_2 = \|\boldsymbol{H}\|_{2,1}\,, \tag{27}$$

where $\boldsymbol{H}_{i,:}$ is the ith row of the abundance matrix $\boldsymbol{H}$.

To conclude, the sparsity term included into the overall loss function is decomposed as

$$\mathcal{L}_{\text{sp}} = \mathcal{L}_{\text{sp-filter}} + \mathcal{L}_{\text{sp-channel}} + \mathcal{L}_{\text{sp-layer}} + \mathcal{L}_{\text{sp-abu}}. \tag{28}$$

Endmember smoothness loss – The endmember perturbation terms are assumed to be spectrally smooth. According to the estimation strategy (9), and the reparametrization (13), the perturbation term affecting the pth endmember in the ith pixel can be identified as $\boldsymbol{v}_{p,i} \in \mathbb{R}^L$ in (13). By gathering all these terms in the 3-dimensional array $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{L\times P\times N}$, the spectral smoothness of these perturbations can be promoted by considering the penalty

$$\mathcal{L}_{\text{smooth}} = \frac{1}{NP}\sum_{\ell=1}^{L-1} \|\boldsymbol{\mathcal{V}}_{\ell,:,:} - \boldsymbol{\mathcal{V}}_{\ell+1,:,:}\|_{\text{F}}^2, \tag{29}$$

To summarize, the loss function of the proposed framework writes

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{data}} + \alpha\mathcal{L}_{\text{KL}} + \beta\mathcal{L}_{\text{sp}} + \gamma\mathcal{L}_{\text{smooth}}, \tag{30}$$

where α, β and γ are positive regularization parameters.

V. Experiments on synthetic images

A. Synthetic data

For a quantitative performance evaluation with respect (w.r.t.) to state-of-the-art unmixing methods, experiments are first conducted on synthetic hyperspectral images. Each generated image is of size 256×256 with 224 spectral bands and is composed of $P = 4$ endmembers. The abundance maps of this set of images are generated using HYDRA toolbox with spatial correlation between local regions [1]. The nominal endmember spectral signatures are extracted from the USGS spectral library, which are captured using a Beckman 5270 spectrometer covering wavelength from 400nm to 2500nm. To introduce physically plausible spectral variability, these signatures are perturbed following the strategy proposed in [44]. More precisely, they are multiplied by piece-wise affine functions, whose one particular example is depicted in Fig. 3. For a amount of variability chosen as $c = 0.15$, such an affine function is built by drawing four parameters, namely, $\xi_i \sim \mathcal{U}(1-c, 1+c)$, $i = \{1, 2, 3\}$ and $L_{\text{break}} \in \{1, \cdots, L\}$. Note a different affine function is generated for each endmember and each pixel. The mixtures are corrupted by a zero mean additive Gaussian noise. The generated data sets are identified with different signal-to-noise ratios, namely, 10dB, 20dB and 30dB.

[1] Available online at http://www.ehu.es/ccwintco/index.php/Hyperspectral Imagery Synthesis tools for MATLAB

TABLE I
RMSE, mSAD and RE results of ablation study using different loss functions. Best results are reported in bold.

Loss functions	**RSME**	**mSAD**	**RE**
$\mathcal{L}_{\text{data}}$	0.0567	0.0742	0.0031
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}}$	0.0478	0.0651	0.0023
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{sp}}$	0.0451	0.0607	0.0019
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{sp}} + \mathcal{L}_{\text{smooth}}$	**0.0424**	**0.0544**	**0.0017**

TABLE II
RMSE, mSAD and RE results using various sparse regularizers and their combinations.

Loss functions	**RMSE**	**mSAD**	**RE**
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{sp-filter}}$	0.0452	0.0592	0.0017
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{sp-channel}}$	0.0465	0.0598	0.0018
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{sp-layer}}$	0.0461	0.0605	0.0018
$\mathcal{L}_{\text{data}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{sp}}$	**0.0424**	**0.0544**	**0.0017**

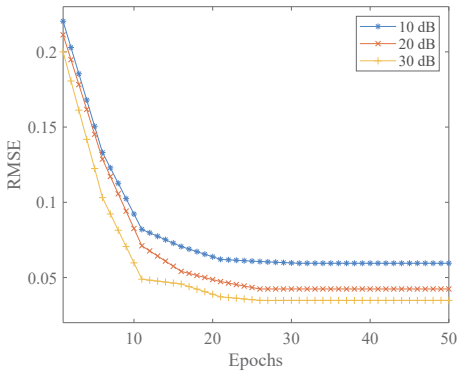

Fig. 4. Synthetic data: RMSE as a function of epochs.

B. Performance figures-of-merit

The unmixing performance is evaluated w.r.t. several figures-of-merit. First, the abundance estimation is assessed

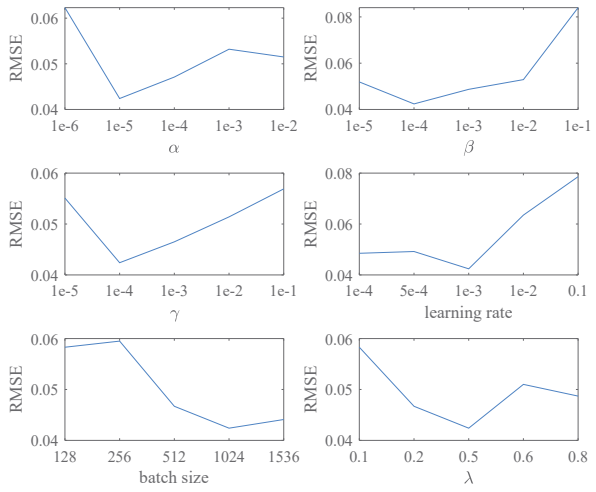

Fig. 5. Synthetic data (SNR = 30dB): RMSE as functions of the parameters α, β, γ, learning rate, batch size and λ.

by the root mean square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{NP}\sum_{i=1}^{N}\|\boldsymbol{a}_i - \hat{\boldsymbol{a}}_i\|^2}, \tag{31}$$

where $\hat{\boldsymbol{a}}_i$ is the estimated abundances, and $\boldsymbol{a}_i$ is the reference abundances. The estimation performance in terms of endmembers is evaluated by computing the mean spectral angle distance (mSAD)

$$\text{mSAD} = \frac{1}{NP}\sum_{i=1}^{N}\sum_{p=1}^{P}\arccos\left(\frac{\langle \boldsymbol{m}_{p,i} \mid \hat{\boldsymbol{m}}_{p,i}\rangle}{\|\boldsymbol{m}_{p,i}\|_2\|\hat{\boldsymbol{m}}_{p,i}\|_2}\right), \tag{32}$$

where $\hat{\boldsymbol{m}}_{p,i}$ and $\boldsymbol{m}_{p,i}$ refer to the true and estimated pth endmember in the ith pixel. Finally, the reconstructed error (RE) is considered to measure data fitting

$$\text{RE} = \frac{1}{NL}\|\boldsymbol{R} - \hat{\boldsymbol{R}}\|_{\text{F}}^2, \tag{33}$$

where $\hat{\boldsymbol{R}}$ are the reconstructed pixels.

C. Ablation analysis

First, an ablation study is conducted to assess the relevance of the different term of the overall loss function (30). The experiments are conducted on a dataset characterized by the SNR chosen as SNR = 20dB. The unmixing results are reported in Table I.

These results show that the sole use of the reconstructed loss in the objective function ensures the completion of the unmixing task, but with a limited accuracy. Including proper regularization can enhance the unmixing performance. Through the introducing of KL loss, the ability of the proposed framework to model the endmember variability is improved. Moreover, sparse constraints are expected to decrease overfitting effects, which may explain the observed corresponding gain. Moreover, promoting sparse abundances takes advantage of a well-documented property inherent to real scenes. The use of the smoothness term guarantees a moderate variation of the endmember variability. To summarize, all terms included in the overall loss function seem to be relevant to reach optimal performances. As an illustration, Fig. 4 presents the RMSE as the number of epochs.

We also conduct an experiment to compare the interest of using the sparsity-promoting regularizations introduced in Section IV-C. The unmixing results are reported in Table II. We observe that all these regularizations lead to significant improvement of the performance, and the combination of all of them provide the best results. Note that the unmixing results of our proposed method are the results of 3DCNN-var/D unless specifically denoted.

D. Comparison with state-of-the-art methods

The performance of the proposed method is compared to those of eight state-of-the-art unmixing methods. The first three methods are recent DNN based unmixing methods using fixed endmembers. The other methods chosen for comparison are designed to handle endmember variability. Their principles and implementation are briefly recalled in what follows.

1) *SDNMF*: the method proposed in [57] is a sparse constrained deep NMF method for unmixing. We set $\lambda = 0.1$, $\mu = 0.1$ and $\alpha = 0.1$ in our work. Abundances and endmembers have been initialized by full constrained least square (FCLS) method [58] and vertex component analysis (VCA) [56] outputs, respectively.
2) *DAEN*: this work [59] is a deep autoencoder based unmixing method. We set $\mu = 0.1$ and $\lambda = 0.1$. VCA is used to initialize the endmembers.
3) *SNMF-Net*: the method proposed in [60] is a model- and learning-based unmixing method. It unrolls the optimization of L_p-sparsity constrained NMF unmixing method into a deep alternating neural network.
4) *ELMM*: the method proposed in [42] considers spectral variability and spatial information. Pixel wise scaling factors are used to take into account the endmember variability. We set the stopping tolerances ϵ_a, ϵ_s and ϵ_Ψ to 1×10^{-3}. We set $\lambda_a = 0.05$, $\lambda_s = 0.001$ and $\lambda_\Psi = 0.001$. FCLS and VCA outputs are applied for initialization.
5) *GELMM*: the method proposed in [43] generalizes ELMM by assuming that the scaling factors are band dependent. The experimental settings of this method are the same as the ELMM.
6) *PLMM*: the method proposed in [44] uses a pure material with an additive perturbation accounting for endmember variability. The tolerance for stopping is set to 1×10^{-3}. We set $\alpha_p = 0.05$, $\beta_p = 2.5 \times 10^{-5}$ and $\gamma = 1$ in our experiments. Abundances and endmembers have been also initialized by FCLS and VCA outputs.
7) *DGEM*: the method in [49] proposes a deep generative endmember model based on VAE to deal with endmember variability. An optimized method is then used to estimate the abundances. The deep generative endmember model is trained with Adam optimizer, and

TABLE III
SYNTHETIC DATA: ESTIMATION PERFORMANCE.

	10 dB			20 dB			30 dB		
	RMSE	**mSAD**	**RE**	**RMSE**	**mSAD**	**RE**	**RMSE**	**mSAD**	**RE**
SDNMF	0.0932	0.4688	0.0824	0.0688	0.4207	0.0481	0.0523	0.3153	0.0238
DAEN	0.0848	0.1855	0.0843	0.0592	0.1559	0.0405	0.0505	0.0701	0.0176
SNMF-Net	0.0914	0.1858	0.1240	0.0861	0.0886	0.0892	0.0589	0.0545	0.0234
ELMM	0.1168	0.1986	0.0303	0.1026	0.1370	0.0102	0.0892	0.1140	0.0080
GELMM	0.1371	0.3633	0.0287	0.0833	0.2759	0.0068	0.0715	0.2714	0.0063
PLMM	0.1716	0.1771	0.0250	0.0846	0.0813	0.0033	0.0495	0.0491	0.0011
DGEM	0.0834	0.0906	0.0229	0.0498	0.0601	0.0029	0.0443	0.0555	0.0014
PGMSU	0.0859	0.0922	0.0335	0.0556	0.0624	0.0193	0.0402	0.0479	0.0073
3DCNN-var/I	0.0604	0.0671	0.0150	0.0433	0.0552	**0.0017**	0.0355	0.0470	0.0002
3DCNN-var/D	**0.0595**	**0.0668**	**0.0149**	**0.0424**	**0.0544**	**0.0017**	**0.0348**	**0.0469**	**0.0002**

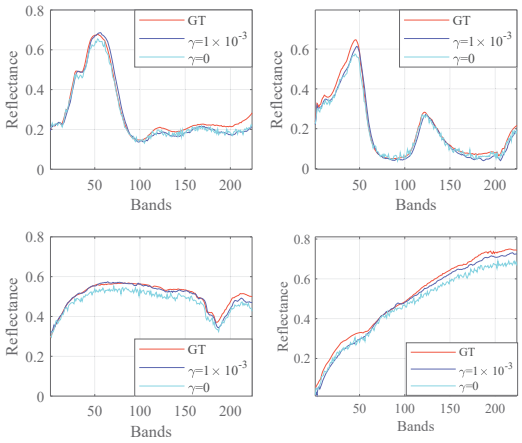

Fig. 6. Synthetic data (SNR = 20dB, one randomly selected pixel): actual endmembers (red) and endmembers estimated with $\gamma = 1 \times 10^{-3}$ (blue) and $\gamma = 0$ (cyan).

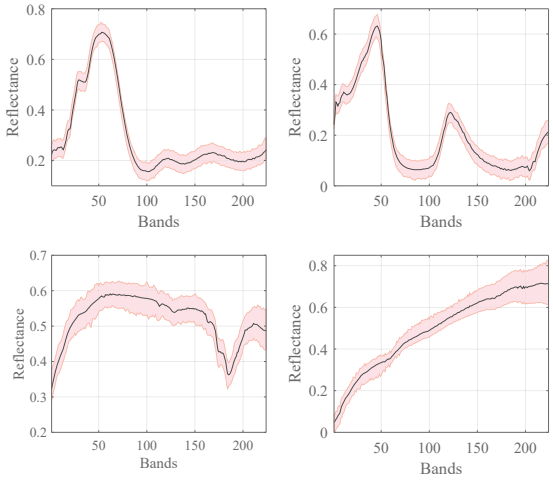

Fig. 7. Synthetic data (SNR = 20dB): estimated nominal endmember signatures (black) and their corresponding estimated variability (red shaded region).

the number of epoch is set to 50. We set $\lambda_A = 0.01$ and $\lambda_Z = 0.1$ in our experiments. Abundances and endmembers have been initialized by FCLS and VCA outputs.

8) *PGMSU*: the work in [61] proposes a VAE-based framework for hyperspectral unmixing accounting for endmember variability. We set $\lambda_1 = 0.1$, $\lambda_2 = 5$, $\lambda_3 = 1$ and $\eta = 0.1$ in our work.
9) The proposed methods (both 3DCNN-var/I and 3DCNN-var/D) have been implemented on Pytorch. Adam optimizer is used to train the framework. The learning rate is set to 1×10^{-3}. The number of training epoch is fixed to 50. The weight decay parameter is set to 1×10^{-5}. The hyperparameters adjusting the weights of the respective terms in the overall loss function have been chosen as $\alpha = 1 \times 10^{-5}$, $\beta = 1 \times 10^{-4}$, $\gamma = 1 \times 10^{-4}$ and $\lambda = 0.5$. Note that the results have been averaged over 10 Monte Carlo runs.

The unmixing results are reported in Table III. The proposed methods provide competitive abundance estimation and endmember extraction results. In particular they achieve good RMSE, aSAD and RE results. These results confirm the effectiveness of the proposed methods, and also indicate that they are more robust to endmember variability. Fig. 8 depicts the abundance maps estimated by the compared methods for the data set corrupted by a noise with SNR = 30dB. After a visual inspection, the abundance maps recovered by the proposed methods seem to be in better agreement with the actual abundance maps. We also observe that 3DCNN-var/D gets better unmixing results than 3DCNN-var/I, which indicates the usefulness of considering correlations across spectral bands. Under the same hardware sources, the running time of 3DCNN-var/I is 390s, and 3DCNN-var/D is 1740s, which shows the trade-off between performance and complexity. Fig. 7 shows the nominal spectral signature as well as the variability of the endmembers estimated by the proposed method (3DCNN-var/D) on the dataset characterized by a noise level of SNR = 20dB. These results are in agreement with an expected spectrally smooth variation around

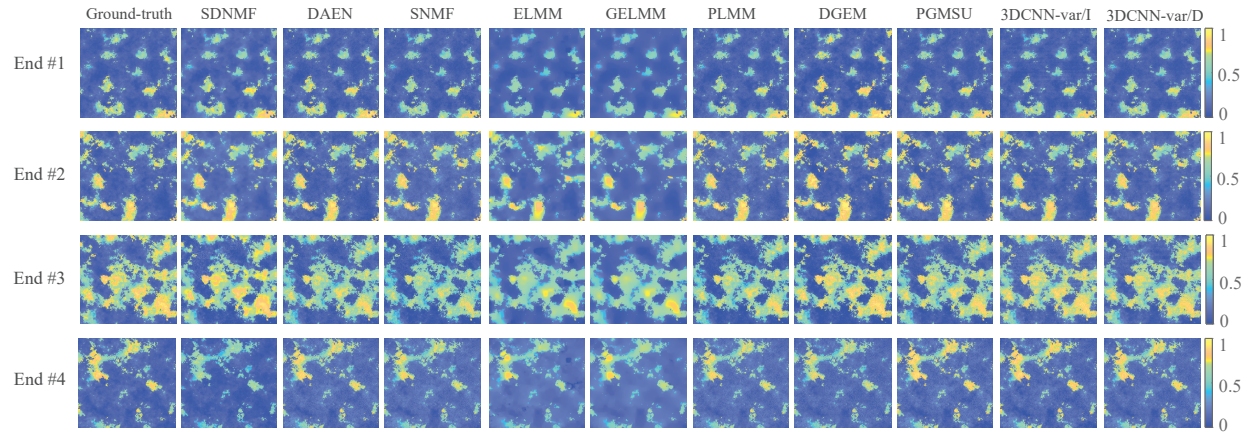

Fig. 8. Synthetic data (SNR = 30dB): actual abundance maps (1st column) and estimated abundance maps (2nd to 11th columns).

TABLE IV
RE COMPARISON OF THE JASPER RIDGE DATA AND URBAN DATA. BEST RESULTS ARE REPORTED IN BOLD.

	Jasper Ridge	Urban
SDNMF	0.0198	0.0432
DAEN	0.0162	0.0185
SNMF-Net	0.0151	0.0172
ELMM	0.0274	0.0122
GELMM	0.0573	0.0121
PLMM	0.0565	0.0118
DGEM	0.0143	0.0112
PGMSU	0.0174	0.0167
3DCNN-var/I	0.0137	0.0105
3DCNN-var/D	**0.0129**	**0.0103**

a nominal signature, demonstrating the ability of the proposed method to handle endmember variability. The parameters of the proposed methods have been selected empirically, using a grid search strategy to choose a set of parameters providing the best unmixing results. Fig. 5 shows the sensitivity of the proposed method (3DCNN-var/D) with respect to α, β, γ, learning rate, batch size and λ. It can be observed the method exhibits satisfactory RMSE within a reasonable range around the optimal parameter values.

VI. EXPERIMENTS ON REAL IMAGES

To further assess the performance and effectiveness of the proposed method, this section reports experiments conducted on real images. Two well-known datasets, namely Jasper Ridge image and Urban image, are considered. The unmixing results of compared algorithms, namely, SDNMF, DAEN, SNMF, ELMM, GELMM, PLMM, DGEM and PGMSU are also presented.

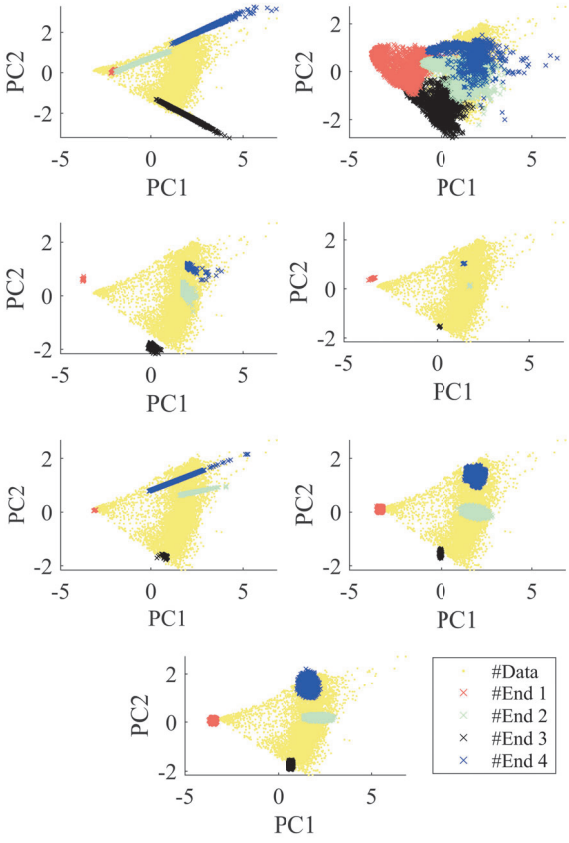

Fig. 9. Jasper image: scatter plot after projection onto the first two principal components. From left to right and top to bottom: ELMM, GELMM, PLMM, DGEM, PGMSU, 3DCNN-var/I and 3DCNN-var/D.

A. Description of the images

The Jasper Ridge image[2] was collected by the airborne visible/infraed imaging spectrometer (AVIRIS) sensor. The scene contains 224 bands covering from 380nm to 2500nm,

[2]Data available online at https://rslab.ut.ac.ir/data.

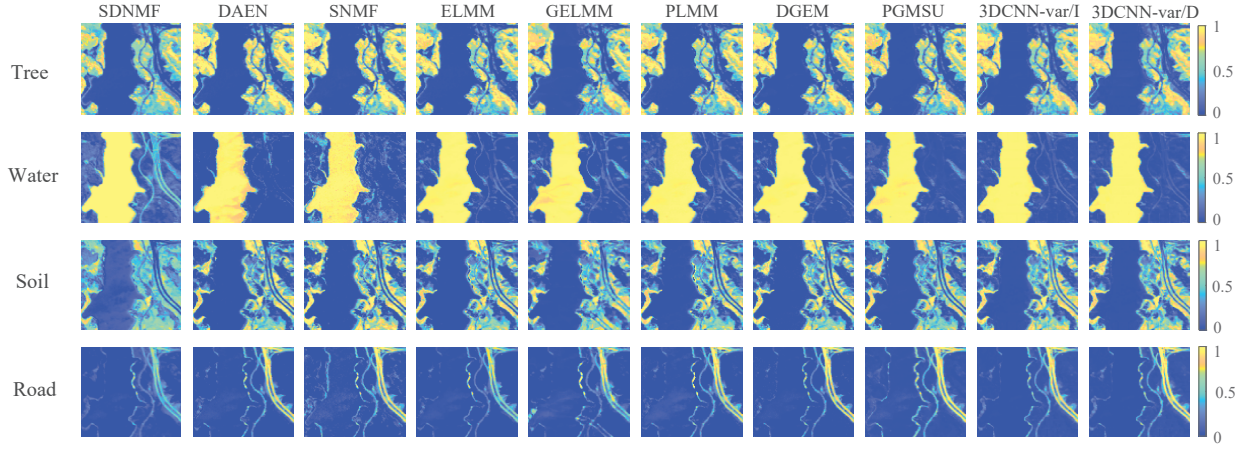

Fig. 10. Jasper image: estimated abundance maps.

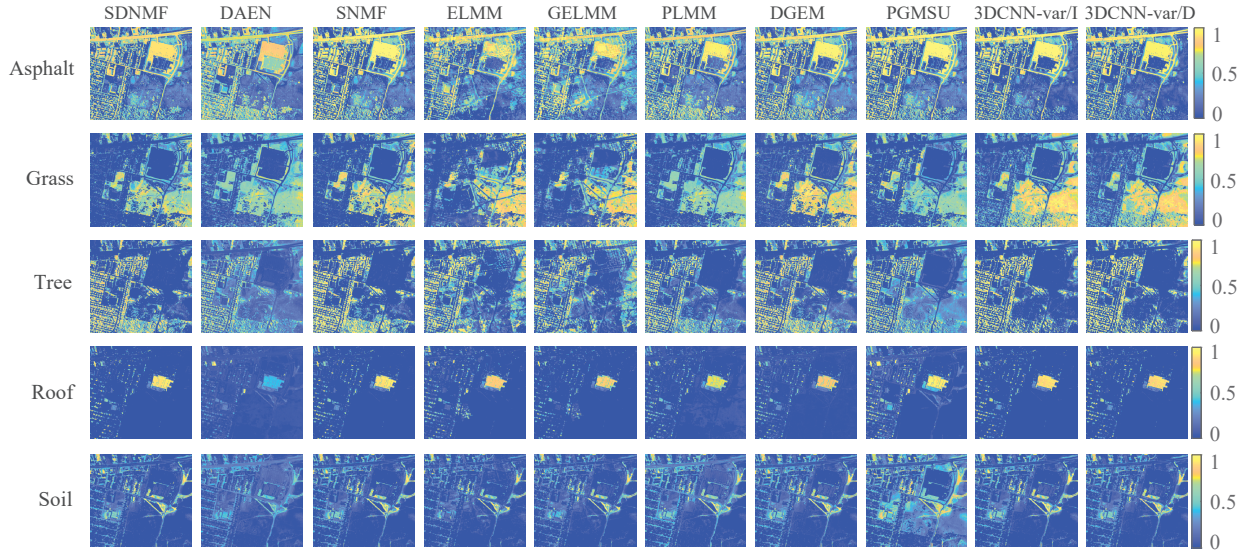

Fig. 11. Urban image: estimated abundance maps.

with a spectral resolution up to 9.46nm. To mitigate the water absorption and atmosphere effects, noisy bands (1–3, 108–112, 154–166 and 220–224) have been removed, leading to $L = 198$ remaining bands. A sub-image of 100×100 pixels has been considered in this experiments. Four main endmembers were identified as "tree", "water", "soil" and "road".

The second scene is the Urban dataset. It was recorded by the hyperspectral digital imagery collection experiment (HYDICE) sensor, and contains 210 bands covering the spectral range $400 - 2400$nm. After removing the noisy and water absorption bands (1–4, 76, 87, 101–111, 136–153 and 198–210), $L = 162$ bands are exploited. The image is composed of an area of 307×307 pixels where each pixel corresponds to a 2m$\times$2m area. Five endmembers have been previously identified as "asphalt", "grass", "tree", "roof" and "soil".

B. Results

The learning rate, training epoch and weight decay parameters used for the proposed method in the real data experiments are the same as those used in synthetic data. In this experiment, α, β, γ and λ are set to 3×10^{-6}, 3×10^{-5}, 5×10^{-3} and 0.1, respectively. Fig. 10 and Fig. 11 show the estimated abundance maps for the Jasper Ridge and Urban datasets, respectively. Compared with other algorithms, the proposed method provides clearer estimated abundance maps with sharper edges.

Fig. 9 depicts the scatter plot of the Jasper dataset after projection onto a plane identified by principal component analysis. The estimated endmembers are also depicted for the compared methods. PLMM and the proposed methods seem to provide the most consistent results. Fig. 12 shows the distribution of the endmembers extracted from the Urban dataset. All these results demonstrate the efficicency of the proposed method

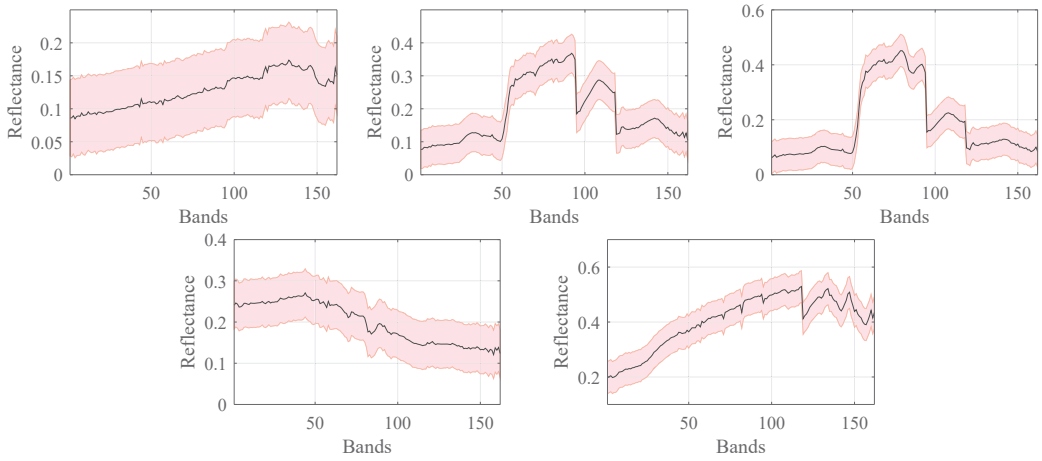

Fig. 12. Urban image: estimated nominal endmember signatures (black) and their corresponding estimated variability (red shaded region).

to cope with endmember variability. Since no ground truth is available to evaluate the estimation performance in terms of abundances and endmembers, only RE is considered as a figure-of-merit. The metrics are reported in Table IV. They show that the proposed method gets the lowest REs.

VII. Conclusion

In this work, we proposed a novel deep autoencoder based framework for blind unmixing, which also accounts for spectral variability. 3D-CNN based architecture was applied to jointly learn spectral-spatial image features, and variational Bayesian learning strategy was applied to approximate the endmember distributions. Structured sparsity-based regularizations were included into the loss function to avoid over-fitting and to promote sparse abundances. Experiments conducted on synthetic and real data show the superior performance of the proposed method when compared to state-of-the-art algorithms. Future works will explore the benefit of more complex deep architectures to model endmember variability.

References

[1] D. W. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Proc. Mag.*, vol. 19, no. 1, pp. 58–69, 2002.

[2] C.-I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, 2004.

[3] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Proc. Mag.*, vol. 19, no. 1, pp. 17–28, 2002.

[4] J. Bolton and P. Gader, "Application of multiple-instance learning for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 889–893, 2011.

[5] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *IEEE Signal Proc. Mag.*, vol. 31, no. 1, pp. 82–94, 2013.

[6] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, 2011.

[7] M. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4484–4502, 2012.

[8] N. Dobigeon, J.-Y. Tourneret, and C.-I. Chang, "Semi-supervised linear spectral unmixing using a hierarchical bayesian model for hyperspectral imagery," *IEEE Trans. Signal Proc.*, vol. 56, no. 7, pp. 2684–2695, 2008.

[9] X. Wang, Y. Zhong, L. Zhang, and Y. Xu, "Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6287–6304, 2017.

[10] J. Chen, C. Richard, and P. Honeine, "Nonlinear estimation of material abundances in hyperspectral images with ℓ_1-norm spatial regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2654–2665, 2013.

[11] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Proc.*, vol. 26, no. 10, pp. 4843–4855, 2017.

[12] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Proc.*, vol. 27, no. 6, pp. 2623–2634, 2018.

[13] K. Makantasis, K. Karantzalos, A. Doulamis, and K. Loupos, "Deep learning-based man-made object detection from hyperspectral data," in *International Symposium on Visual Computing*. Springer, 2015, pp. 717–727.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[15] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, 2019.

[16] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised Spatial–Spectral Feature Learning by 3D Convolutional Autoencoder for Hyperspectral Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, 2019.

[17] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, 2014.

[18] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Proc. Mag.*, vol. 31, no. 1, pp. 95–104, 2013.

[19] Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Bilinear models for nonlinear unmixing of hyperspectral images," in *Proc. IEEE GRSS Workshop on Hyperspectral Image and SIgnal Processing: Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal, June 2011, pp. 1–4.

[20] B. Hapke, "Bidirectional reflectance spectroscopy: 1. Theory," *Journal of Geophysical Research: Solid Earth*, vol. 86, no. B4, pp. 3039–3054, 1981.

[21] R. Ammanouil, A. Ferrari, C. Richard, and S. Mathieu, "Nonlinear unmixing of hyperspectral data with vector-valued kernel functions," *IEEE Trans. Image Proc.*, vol. 26, no. 1, pp. 340–354, 2016.

[22] B. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Hyperspectral unmixing using a neural network autoencoder," *IEEE Access*, vol. 6, pp. 25 646–25 656, 2018.

[23] R. Guo, W. Wang, and H. Qi, "Hyperspectral image unmixing using autoencoder cascade," in *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2015, pp. 1–4.

[24] Y. Qu and H. Qi, "uDAS: An untied denoising autoencoder with sparsity for spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1698–1712, 2018.

[25] Y. Su, A. Marinoni, J. Li, A. Plaza, and P. Gamba, "Nonnegative sparse autoencoder for robust endmember extraction from remotely sensed hyperspectral images," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 205–208.

[26] S. Ozkan and G. B. Akar, "Deep spectral convolution network for hyperspectral unmixing," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3313–3317.

[27] Y. Su, A. Marinoni, J. Li, J. Plaza, and P. Gamba, "Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1427–1431, 2018.

[28] Y. Su, X. Xu, J. Li, H. Qi, P. Gamba, and A. Plaza, "Deep autoencoders with multitask learning for bilinear hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[29] M. Wang, M. Zhao, J. Chen, and S. Rahardja, "Nonlinear unmixing of hyperspectral data via deep autoencoder networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1467–1471, 2019.

[30] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral-spatial hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[31] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[32] K. Wei, Y. Fu, and H. Huang, "3-D quasi-recurrent neural network for hyperspectral image denoising," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 363–375, 2020.

[33] O. Sidorov and J. Yngve Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[34] F. Khajehrayeni and H. Ghassemian, "Hyperspectral unmixing using deep convolutional autoencoders in a supervised scenario," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 13, pp. 567–576, 2020.

[35] C. Bateson, G. Asner, and C. Wessman, "Endmember bundles: a new approach to incorporating endmember variability into spectral mixture analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 1083–1094, Mar 2000.

[36] B. Somers, M. Zortea, A. Plaza, and G. P. Asner, "Automated extraction of image-based endmember bundles for improved spectral unmixing," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 5, no. 2, pp. 396–408, 2012.

[37] M. A. Veganzones, G. Tochon, M. Dalla-Mura, A. J. Plaza, and J. Chanussot, "Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation," *IEEE Trans. Image Proc.*, vol. 23, no. 8, pp. 3574–3589, 2014.

[38] T. Uezato, R. J. Murphy, A. Melkumyan, and A. Chlingaryan, "A novel endmember bundle extraction and clustering approach for capturing spectral variability within endmember classes," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6712–6731, 2016.

[39] D. A. Roberts, M. Gardner, R. Church, S. Ustin, G. Scheer, and R. Green, "Mapping chaparral in the santa monica mountains using multiple endmember spectral mixture models," *Remote sensing of environment*, vol. 65, no. 3, pp. 267–279, 1998.

[40] F. Chen, K. Wang, and T. F. Tang, "Spectral unmixing using a sparse multiple-endmember spectral mixture model," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5846–5861, 2016.

[41] T. Uezato, M. Fauvel, and N. Dobigeon, "Hyperspectral unmixing with spectral variability using adaptive bundles and double sparsity," *IEEE Trans. Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3980–3992, June 2019.

[42] L. Drumetz, M.-A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Trans. Image Proc.*, vol. 25, no. 8, pp. 3890–3905, 2016.

[43] T. Imbiriba, R. A. Borsoi, and J. C. M. Bermudez, "Generalized linear mixing model accounting for endmember variability," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1862–1866.

[44] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Trans. Singal Proc.*, vol. 64, no. 2, pp. 525–538, 2015.

[45] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. application to hyperspectral imagery," *IEEE Trans. Image Proc.*, vol. 19, no. 6, pp. 1403–1413, 2010.

[46] X. Du, A. Zare, P. Gader, and D. Dranishnikov, "Spatial and spectral unmixing using the beta compositional model," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 7, no. 6, pp. 1994–2003, 2014.

[47] Y. Zhou, A. Rangarajan, and P. D. Gader, "A Gaussian mixture model representation of endmember variability in hyperspectral unmixing," *IEEE Trans. Image Proc.*, vol. 27, no. 5, pp. 2242–2256, 2018.

[48] D. Hong, L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, and B. Zhang, "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[49] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Deep generative endmember modeling: An application to unsupervised spectral unmixing," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 374–384, 2019.

[50] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Online unmixing of multitemporal hyperspectral images accounting for spectral variability," *IEEE Trans. Image Proc.*, vol. 25, no. 9, pp. 3979–3990, 2016.

[51] J. Sigurdsson, M. O. Ulfarsson, J. R. Sveinsson, and J. M. Bioucas-Dias, "Sparse distributed multitemporal hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6069–6084, 2017.

[52] J. M. Bioucas-Dias and J. M. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, 2008.

[53] M. Zhao, M. Wang, J. Chen, and S. Rahardja, "Hyperspectral unmixing for additive nonlinear models with a 3-d-cnn autoencoder network," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[54] T. Muschinski, G. J. Mayr, T. Simon, and A. Zeileis, "Cholesky-based multivariate gaussian regression," *arXiv preprint arXiv:2102.13518*, 2021.

[55] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *arXiv preprint arXiv:1906.02691*, 2019.

[56] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, 2005.

[57] X.-R. Feng, H.-C. Li, J. Li, Q. Du, A. Plaza, and W. J. Emery, "Hyperspectral unmixing using sparsity-constrained deep nonnegative matrix factorization with total variation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6245–6257, 2018.

[58] D. C. Heinz and C. Chang, "Fully constrained least-squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 29, no. 3, pp. 529–545, Mar. 2001.

[59] Y. Su, J. Li, A. Plaza, A. Marinoni, P. Gamba, and S. Chakravortty, "DAEN: Deep autoencoder networks for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4309–4321, 2019.

[60] F. Xiong, J. Zhou, S. Tao, J. Lu, and Y. Qian, "SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[61] S. Shi, M. Zhao, L. Zhang, and J. Chen, "Variational autoencoders for hyperspectral unmixing with endmember variability," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1875–1879.

Min Zhao (Graduate Student Member, IEEE) received the B.S. degree in electronic and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2017, and the M.S. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, in 2020, where she is pursuing the Ph.D. degree. Her research interests include hyperspectral image analysis and object detection. She received the People's choice Award of Three Minute Thesis (3MT) Competition of IEEE IGARSS 2020, and the (with team) Champion of Grand Challenges on NIR Image Colorization of IEEE VCIP 2020.

Shuaikai Shi (Graduate Student Member, IEEE) received the B.S. degree in applied physics from Taiyuan Normal University, Taiyuan, China, in 2016, and the M.S. degree in physics from Tongji University, Shanghai, in 2019. Now he is pursuing the Ph.D. degree with Northwestern Polytechnical University. His current research interests are Bayesian machine learning and its applications in hyperspectral image processing, particularly, in data unmixing and fusion.

Jie Chen (Senior Member, IEEE) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2006, the Dipl.-Ing. and M.S. degrees in information and telecommunication engineering from the University of Technology of Troyes (UTT), Troyes, France, and Xi'an Jiaotong University, respectively, in 2009, and the Ph.D. degree in systems optimization and security from UTT in 2013.

From 2013 to 2014, he was with the Lagrange-Laboratory, University of Nice Sophia Antipolis, Nice, France. From 2014 to 2015, he was with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. He is a Professor with Northwestern Polytechnical University, Xi'an, China. His research interests include adaptive signal processing, distributed optimization, hyperspectral image analysis, and acoustic signal processing.

Dr. Chen was the Technical Co-Chair of IWAENC'16 held in Xi'an, China. He serves as a Distinguished Lecturer for Asia–Pacific Signal and Information Processing Association (2018–2019), and a Co-Chair of the IEEE Signal Processing Society Summer School 2019 in Xi'an. He will be the General Co-Chair of IEEE MLSP 2022.

Nicolas Dobigeon (S'05–M'08–SM'13) received the Eng. degree in electrical engineering from ENSEEIHT, Toulouse, France, and the M.Sc. degree in signal processing from Toulouse INP, both in 2004, and the Ph.D. degree and the Habilitation à Diriger des Recherches in signal processing from Toulouse INP in 2007 and 2012, respectively.

From 2007 to 2008, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. Since 2008, he has been with INP-ENSEEIHT Toulouse, University of Toulouse, where he is currently a Professor. He conducts his research within the Signal and Communications Group, IRIT Laboratory, and he currently holds an AI Research Chair at the Artificial and Natural Intelligence Toulouse Institute (ANITI). His recent research activities have been focused on statistical signal and image processing, with a particular interest in Bayesian inverse problems and applications to remote sensing, biomedical imaging, astrophysics and microscopy. He is a Junior Member of Institut Universitaire de France (IUF).