5-2018

Hyperspectral Image Analysis through Unsupervised Deep Learning

Ying Qu
University of Tennessee, yqu3@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

To the Graduate Council:

I am submitting herewith a dissertation written by Ying Qu entitled "Hyperspectral Image Analysis through Unsupervised Deep Learning." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Engineering.

Hairong Qi, Major Professor

We have read this dissertation and recommend its acceptance:

Jens Gregor, Lynne E. Parker, Jindong Tan

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Hyperspectral Image Analysis through Unsupervised Deep Learning

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Ying Qu

May 2018

Dedication

This dissertation is dedicated to my parents, my brother and my husband.

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Hairong Qi. Dr. Qi accepted me to the group in November, 2015. Her persistence and patience always calm me down when I face pressures and challenges. Dr. Qi is a role model, who cares about her students in both research and life. Her insights, broad knowledge and open minds guide me to a better way when I reach a dead end in research. Without her support and guidance, I would not have been able to develop my research skills and complete this dissertation work.

I also would like to thank Dr. Lynne E. Parker, who always inspires me by her lectures and talks. Her insights for careers and life have been very instructive and important. I am also grateful to Dr. Jens Gregor, for the time and useful advices in my research. My deep appreciation also goes to Dr. Jingdong Tan, who provided useful suggestions when I felt frustrated. I greatly appreciate my committee members for their time and input to improve this dissertation.

I learned a lot from our group members since I joined the group, especially their attitude and skills on research. I would further thank all my labmates and alumni in the AICIP group, for their valuable help and friendship.

Last but not the least, I would like to express my deepest appreciation to my parents, my brother, and my husband for their unconditional love and encouragement.

Abstract

Hyperspectral image (HSI) analysis has become an active research area in computer vision field with a wide range of applications. However, in order to yield better recognition and analysis results, we need to address two challenging issues of HSI, *i.e.*, the existence of mixed pixels and its significantly low spatial resolution (LR). In this dissertation, spectral unmixing (SU) and hyperspectral image super-resolution (HSI-SR) approaches are developed to address these two issues with advanced deep learning models in an unsupervised fashion. A specific application, anomaly detection, is also studied, to show the importance of SU.

Although deep learning has achieved the state-of-the-art performance on supervised problems, its practice on unsupervised problems has not been fully developed. To address the problem of SU, an untied denoising autoencoder is proposed to decompose the HSI into endmembers and abundances with non-negative and abundance sum-to-one constraints. The denoising capacity is incorporated into the network with a sparsity constraint to boost the performance of endmember extraction and abundance estimation.

Moreover, the first attempt is made to solve the problem of HSI-SR using an unsupervised encoder-decoder architecture by fusing the LR HSI with the high-resolution multispectral image (MSI). The architecture is composed of two encoder-decoder networks, coupled through a shared decoder, to preserve the rich spectral information from the HSI network. It encourages the representations from both modalities to follow a sparse Dirichlet distribution which naturally incorporates the two physical constraints of HSI and MSI. And the angular difference between representations are minimized to reduce the spectral distortion.

Finally, a novel detection algorithm is proposed through spectral unmixing and dictionary based low-rank decomposition, where the dictionary is constructed with mean-shift clustering and the coefficients of the dictionary is encouraged to be low-rank. Experimental evaluations

show significant improvement on the performance of anomaly detection conducted on the abundances (through SU).

The effectiveness of the proposed approaches has been evaluated thoroughly by extensive experiments, to achieve the state-of-the-art results.

Table of Contents

List of Tables

List of Figures

xiii

Chapter 1

Introduction

1.1 Problem Statement

Compared to multispectral image (MSI with around 10 spectral bands) or conventional color image (RGB with 3 bands), hyperspectral image (HSI) collects hundreds of contiguous bands which provide finer details of spectral signature of different materials, that extended beyond the visible bands, *e.g.*, near infrared bands [18]. An example of HSI is illustrated in Fig. 1.1. Each pixel in the HSI consists of reflection intensities at hundreds of bands, acquired by a hyperspectral camera with an interval of 10-20 nm.
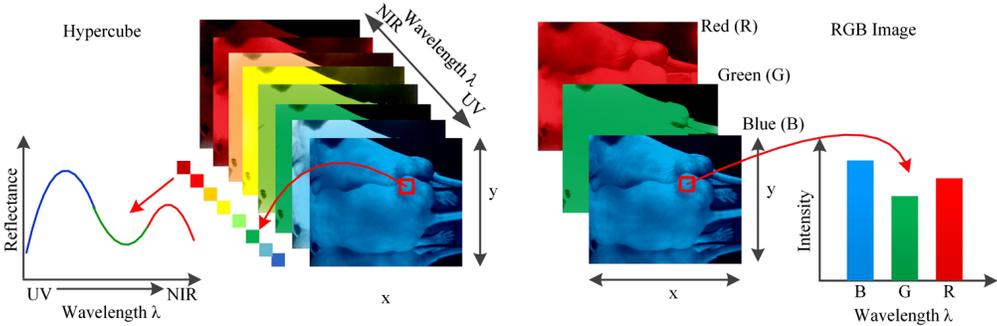
Figure 1.1: Hyperspectral image with multiple bands and conventional image with three channels [101].

HSI analysis has become a thriving and active research topic in computer vision and remote sensing fields with a wide range of applications [22, 17], such as object recognition and classification [84, 49, 169, 103], tracking [149, 50, 146, 147], environmental monitoring [139,

121] and change detection [85, 18]. However, to fully take advantage of HSI, there are two challenging issues need to be addressed, *i.e.*, the existence of mixed pixels and its significant low spatial resolution.

HSI has a common issue in terms of the existence of mixed pixels, i.e., the measured spectrum of a single pixel is a mixture of several constituent spectrum, or endmembers, weighted with corresponding fractional coefficients, or abundances. Very often, one need to decompose the HSI data into endmembers and abundance before analysis. This procedure is also referred to as spectral unmixing (SU). In general, spectral unmixing involves three major steps: the number of endmembers estimation, endmembers extraction and abundance estimation. Since the decomposed endmember and abundance matrices represent materials and fractional coefficients, respectively, both matrices subject to the non-negative constraint. In particular, each column of the abundance should be sum-to-one. To solve the unmixing problem, most methods rely on the assumption of linear mixing model [17, 79, 6], i.e., different spectra do not interfere with each other, as illustrated in Fig. 1.2.
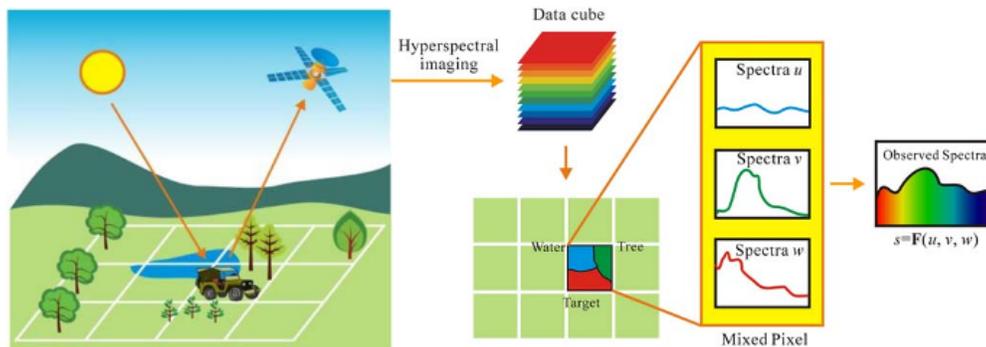
Figure 1.2: Hyperspectral image linear mixing model. [171]

Besides the problem of mixed pixels, another crucial issue is that the spatial resolution of HSI is significantly lower than conventional images. To acquire high spectral resolution, the spectrum has to be divided into many bands; thus in each band, only a small amount of energy can reach the sensor, making it impractical to capture images with high spatial resolution. On the contrary, the conventional images, *e.g.*, RGB, could record total intensity of radiation falling on the sensors; therefore presenting high spatial resolution although with poor spectral resolution (e.g., only three bands). In real applications, images with both high spectral and spatial resolution are desired [152] to yield better recognition and

analysis results. A natural way to generate such images is to fuse hyperspectral image with multispectral image or conventional color image. This procedure is referred to as *hyperspectral image super-resolution (HSI-SR)* [4, 87, 37] or *hyperspectral pansharpening.*

As an application of HSI, anomaly detection (AD) has become one of the hottest topic in HSI analysis over the last 20 years [120, 108, 16, 174]. It aims to identify unusual components in the HSI data. As shown in Fig. 1.3, AD is an unsupervised classification problem where the anomaly is assumed to be sparse and the background is widely populated [108]. The performance of AD would be improved by solving the problem of SU.

Nowadays, deep learning (DL) has been rapidly developed and extended for HSI analysis with tremendous accomplishment in different applications, *e.g.*, feature extraction and classification [27, 125, 175, 78, 169, 103], semantic annotation [163, 133, 153], object detection [60, 32, 143, 167, 31], segmentation [76], change detection [53, 176, 80] and data fusion [172, 70, 21]. DL is a specific class of machine-learning approaches, and it is able to solve complex tasks that are difficult for traditional machine-leaning methods. DL can be traced back to 1943, when Walter Pitts and Warren McCulloch created a computer model based on neural networks to simulate human brain [155]. In the early 90's, LeCun *et al.* [90] introduced the first convolutional neural network (CNN, an architecture of DL) to recognize visual patterns with a hierarchical structure. When the CNN structure conquered more complicated AI tasks [155] in 2012, the DL begins to bloom.
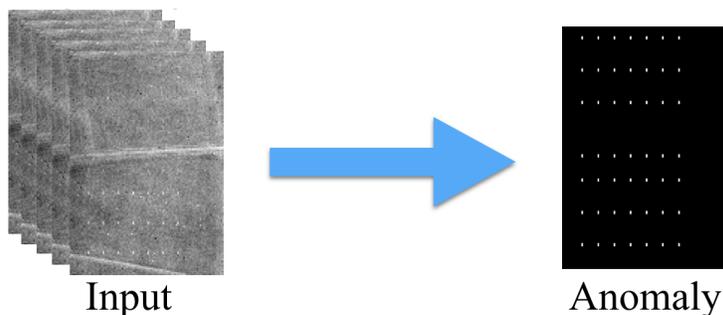
Figure 1.3: Hyperspectral image anomaly detection, the output anomaly is sparse pixels in the images.

In this dissertation, our objective is to explore the possibility of solving SU and HSI-SR with advanced deep learning approaches. In addition, the challenge anomaly detection problem is also studied as a HSI application.

1.2 Motivations

As described in Sec 1.1, HSI has been utilized in various applications in computer vision and remote sensing fields, because it provides rich spectral information that can distinguish different materials. However, there are two challenging issues associated with HSI, which may hinder its widespread adoption, *i.e.*, the existence of mixed pixels and its extremely low spatial resolution. Since the ground truth is not available, these two issues need to be addressed in an unsupervised fashion.

First, due to the large footprint, the intensity at each pixel of HSI is a mixture of reflectance from several constituent pure materials covered within the pixel. Thus spectral unmixing (SU) is a preliminary step in many applications [79, 17] in order to fully take advantages of HSI. In the past few decades, there have been numeral traditional approaches developed to solve the unmixing problem. However, for most approaches, noise remains a challenge which would dramatically decrease the unmixing accuracy, especially when the number of endmembers is wrong. From another perspective, the unmixing procedure can also be explained as finding a set of low-dimensional representations (i.e., abundance) that reconstruct the data with their corresponding dictionaries (i.e., endmembers), which matches the purpose of autoencoder. Here, a connection between spectral unmixing and autoencoder is established . The challenges to address this problem are how to incorporate two fundamental requirements, *i.e.*, non-negative and sum-to-one, as well as noise handling.

Second, due to the hardware limitations, one can only expect to acquire images of high resolution in either the spatial or spectral domains. Although the spectral resolution of HSI is high, its spatial resolution is significantly lower than conventional images, *i.e.*, RGB, MSI. On the other hand, conventional images carry much higher spatial resolution despite of its low spectral resolution. In order to obtain images with high resolution in both domains, hyperspectral super-resolution (HSI-SR) methods are demanded to fuse these two modalities. Existing HSI-SR approaches [100, 179, 87, 165] generally assume that the LR HSI is downsampled from the HR HSI and such down-sampling function is used as a prior in the fusion methods, which may not be true in practice due to the distortions caused by both the sensors and complex environmental conditions [4]. HSI-SR is closely related to the natural image

super-resolution (SR) problem, which has been extensively studied and achieved excellent performance through the state-of-the-art *deep learning* [42, 102, 132, 81, 82, 91, 86, 62]. However, these deep learning based methods are all supervised and require a large dataset, the down-sampling function, and the availability of the ground truth HR HSI, making their adoption on the problem of HSI-SR a challenge. Therefore, I would like to make the first effort to solve this problem with unsupervised deep learning.

Third, an unsupervised anomaly detection problem is further explored, which benefits from solving the problem of SU. Anomaly detection is essential to many real world problems, such as environmental monitoring and mineral reconnaissance. The anomaly detection can be modeled as an unsupervised binary classification problem between the background class and the anomaly class. The challenge of this problem is that there is no prior knowledge of the anomalies or the background. The background could have complex textures which will increase the difficulty of detection. Also due to the limitations of HSI acquisition devices, HSI data are usually corrupted by noise. Fortunately, there are two important characteristics that could be used to distinguish anomalies from their background, i.e., compared to other objects, the anomalies are 1) sparse, and 2) possess distinctive features as compared to their surrounding background. Therefore, in order to improve the performance of anomaly detection, how to take this two properties into the objective function is the key to address this problem.

Although deep learning has achieved the state-of-the-art performance on supervised problems, its practice on unsupervised problems has not been fully developed. In this dissertation, both problems, SU and HSI-SR, are explored through the advanced unsupervised deep learning. In addition, an unsupervised classification problem, *i.e.*, anomaly detection is studied which benefits from addressing these two issues. The whole flowchart is shown in Fig. 1.4.
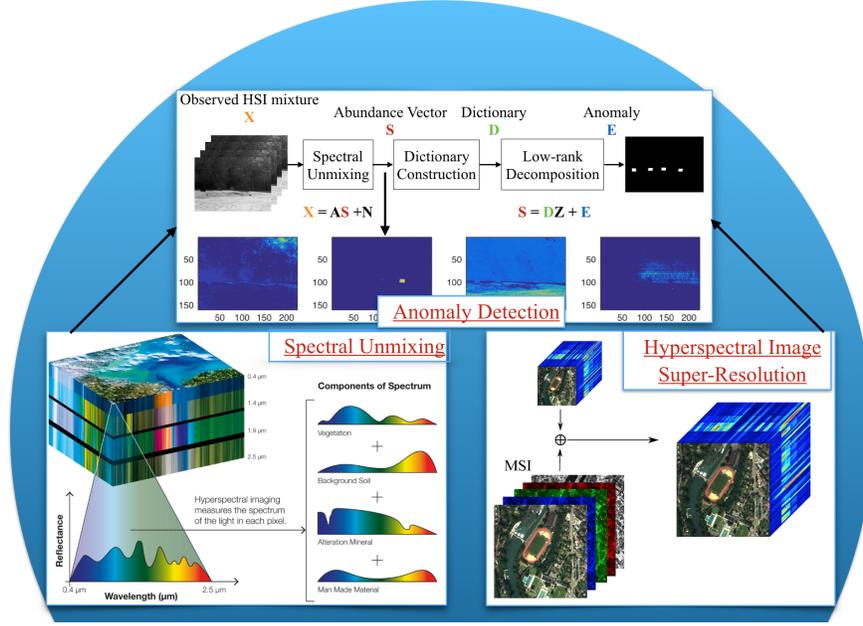
Figure 1.4: Hyperspectral images analysis including spectral unmixing, hyperspectral image super-resolution and anomaly detection.

1.3 Contributions

In this dissertation, three novel approaches are developed to solve the aforementioned problems, *i.e.*, spectral unmixing, hyperspectral image super-resolution and anomaly detection.

The contribution of spectral unmixing is three-fold. First, through proof by counterexample, we make important discovery that existing part-based autoencoders with non-negative tied encoder and decoder would not be able to extract the accurate endmembers. Based on this discovery, we propose an innovative autoencoder network for spectral unmixing purpose where the encoder and decoder are independent and only the decoder is enforced to be non-negative. Second, traditional approaches usually attach a denoiser in front of the unmixing procedure, which would introduce additional denoising error before unmixing. In order to reduce such error, we propose an end-to-end autoencoder network where the denoising capacity is incorporated as a constraint to the network, thus avoiding the introduction of extra error to the network. Third, the unmixing performance drops drastically when there exists noise and the number of endmembers is incorrectly estimated. To address this problem,

an l_{21} constraint is applied to reduce the redundant rows of the encoder, which improves the performance of abundance estimation.

An *unsupervised* network structure is proposed to address the challenges of HSI-SR. To the best of our knowledge, this is the first effort to solving the HSI-SR problem with deep learning in an unsupervised fashion. The novelty of this work is three-fold. First, the network extracts both the spectral and spatial information from LR HSI and HR MSI with two deep learning networks which share the same decoder weights, as illustrated in Fig. 4.2. Second, in order to incorporate the two physical constraints of HSI and MSI data representation, *i.e.*, sum-to-one and sparsity, the network encourages the representations from both modalities to follow a Dirichlet distribution which naturally incorporates the sum-to-one property. Since each pixel of the image only consists of a few spectral bases, the sparsity of the representations is guaranteed by minimizing their entropy function. Third, to address the problem of spectral distortion, instead of adopting the down-sampling function (as an estimated mapping function) to relate the representations of both modalities, the angular difference of these representations is minimized so that they have similar patterns. In this way, the spectral distortion is largely reduced. The proposed method is referred to as uSDN.

To overcome the drawbacks of existing approaches, an anomaly detection algorithm is proposed based on spectral unmixing and low-rank decomposition. The uniqueness of the proposed approach can be summarized from two perspectives. *First*, since HSI images are normally highly mixed, for better anomaly detection performance, the HSI pixels are projected into a single subspace, where anomalies should be distinct from the background. There are many options that can be adopted for this purpose, *e.g.*, principal component analysis (PCA) [33], sparse-representation [28], etc. However, in order to preserve the physical meaning of the data, the unmixing method is adopted, which is a part-based representation allowing only additive combinations. Compared to other representations, the non-negativity constraint in the unmixing approaches aligns well with intuition: the whole is equal to the sum of parts [92]. Therefore, instead of using the raw pixels to detect anomalies, the spectral unmixing algorithm is applied to obtain the abundance vectors. The unmixing serves the purpose of removing noise. But more importantly, the abundance vectors possess

more distinctive patterns in discriminating anomalies from the background. *Second*, since the background is highly correlated, a dictionary is constructed to better describe the background and sparse anomalies where the dictionary atoms are determined using results from the non-parametric mean-shift clustering algorithm. In this way, instead of directly applying existing low-rank decomposition approaches, a dictionary-based low-rank decomposition algorithm is proposed according to the properties of the background and anomaly. The proposed approach is able to extract the sparse anomalies where the low-rank constraint is applied on the background coefficients, instead of the background itself, such that the noisy pixels in the anomaly matrix can be removed to a large extent. The proposed anomaly detection method is referred to as abundance- and dictionary-based low-rank decomposition, or ADLR.

1.4   Dissertation Organization

The dissertation is organized as follows: Chapter 1 serves as an introduction which describes the studied problems, as well as the major contributions of this dissertation. Chapter 2 presents a literature review on our studied problems. Chapter 3 and chapter 4 describe the proposed unsupervised spectral unmixing and hyperspectral image super-resolution approaches developed with deep learning. Chapter 5 proposes the anomaly detection approach through spectral unmixing and low-rank decomposition. In chapters 3-5, the proposed approaches are evaluated with comprehensive experiments, and compared with the state-of-the-art methods. Finally, the summary and future work of this dissertation are discussed in Chapter 6.

Chapter 2

Literature Review

In this chapter, we study the literatures that covering the problems needed to be addressed. We start with an explanation of the fundamental knowledge of deep learning, which related to our proposed approaches. We then study the state-of-the-art approaches that are widely used to solve the spectral unmixing, hyperspectral image super-resolution and anomaly detention problems, respectively.

2.1   Deep Learning

Deep learning has been extensively studied and achieved incredible success in computer vision field [89, 54, 155]. The widely-used four typical structures in deep learning include deep belief network (DBN), autoencoder (AE), convolution neural network (CNN) and recurrent neural network (RNN). The former two structures belong to the unsupervised learning, which have the potential to solve our problems. DBN was initially developed by Hinton [65]. Since it adopted the restricted Boltzmann machines (RBM), the training procedure requires extensive computation. Therefore, in this dissertation, autoencoder is studied as a potential method to solve our problems.

2.1.1 Autoencoder

Autoencoder (AE) was first proposed in 1980s by Hinton's group [126]. This architecture learns the representations by reducing the reconstruction error of the given data. It consists of encoder and decoder layers, as shown in Fig. 2.1.

Given the input data x, the hidden layer (representations) $z = f(x) = \sigma(\mathbf{W}_1 x + b_1)$, where f denotes the encoder of the network. Similarly, the reconstructed data $\hat{x} = g(z) = \sigma(\mathbf{W}_2 z + b_2)$, where g denotes the decoder of the network. Note that σ denotes the activation function of each node. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices of encoder and decoder, and b_1, b_2 are the bias vectors.
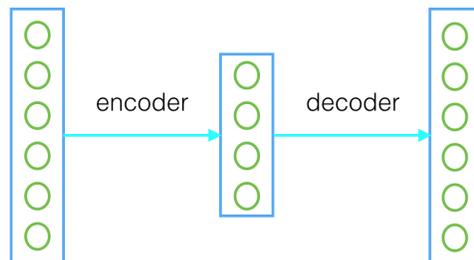
Figure 2.1: Autoencoder architecture.

To obtain more meaningful representations, denoising autoencoder (DA) [151] was developed to learn representations from noisy data. The DA is trained to reconstruct the clean data x from the randomly corrupted data $\hat{x}$. Assuming that there exists a linear transformation function between the reconstructed clean data x and the noisy data $\hat{x}$, the marginalized denoising autoencoder (MDA) [26] improves the traditional DA by marginalizing the corruption. The clean data can then be evaluated using a close-form solution only with respect to the input noisy data. To achieve sparse representations, the sparse autoencoder [115] was developed by introducing additional cost function to reduce the active values of the hidden layer. Variational autoencoder [40] adopted strong assumptions of hidden layer by introducing stochastic variational inference.

2.1.2 Multi-Modal Deep Learning

HSI-SR is to fuse multiple input sources by finding shared representations, which can be considered as multi-modal learning. Multi-modal learning refers to relating multiple modalities together with shared information [116], even different modalities may have very distinct statistical properties. In 2011, Ngiam *et al.* [116] proposed a deep multi-modal autoencoder to learn both the audio and visual bimodal features from the data. Given one source, the network is trained to reconstruct both sources by learning representations shared by audio and visual. It is initialized with the bimodal DBN weights [66] to prevent local minimum. In particular, the author demonstrated that it is able to learn more robust representations for one modality if multi-modalities are offered for representation learning. Srivastava and Salakhutdinov [140] proposed an deep-Boltzmann machine based multi-modle learning, which learned a generative model given multi-modal data. This generative model can handle modalities with missing or noisy information effectively. In 2015, Eitel *et al.* [48] proposed a novel RGB-D multi-modal architecture for object recognition through Constitutional Neural Networks (CNN). Two CNNs are built for RGB image and depth image, respectively, to learn a shared representation for object recognition.

2.2 Spectral Unmixing

2.2.1 Linear Mixture Model

In most cases, if neglecting the interference of different materials, the observed pixel can be well described using a linear mixing model (LMM) [17]. Under this model, each pixel in of HSI is assumed to be a linear combination of spectra (endmembers) $\mathbf{A}$, weighted with their corresponding fractions (abundance) $\mathbf{S}$.

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N} \tag{2.1}$$

where $\mathbf{X} = \{x_1, x_2, ...x_n\} \in \mathbb{R}^{l \times n}$ is the observed mixture with l spectral bands and n data samples, $\mathbf{A} = \{a_1, a_2, ...a_c\} \in \mathbb{R}^{l \times c}$ denotes c endmembers, and $\mathbf{S} = \{s_1, s_2, ...s_n\} \in \mathbb{R}^{c \times n}$

denotes their abundance. The matrix $\mathbf{N} \in \mathbb{R}^{l \times n}$ denotes the noise. At a given pixel, the fractional abundance subject to the non-negative and sum-to-one constrains.

$$s_{ij} \geq 0 \quad \text{and} \quad \sum_{i=1}^{c} s_{ij} = 1; \tag{2.2}$$

In the past few decades, a number of methods have been proposed to tackle the unmixing problem and most approaches can be categorized as geometrical based and statistic based approaches.

2.2.2 Geometrical based Approaches

Geometrical-based approach is one of the most popular branches for linear unmixing. The concept is illustrated in Fig. 2.2. Under the assumption that there exists a convex simplex which circumscribes the observed data, the geometrical based approaches aim to find the vertices of the simplex as the endmembers of the given data. Widely used methods like N-FINDER [159] and vertex component analysis (VCA) [113] belong to this category. N-FINDER [159] is based on the fact that the largest volume should be built by the purest pixels in the data. Thus the set of endmembers is found by inflating the simplex inside the data. VCA iteratively projects data onto the direction that is orthogonal to the subspace spanned by the already determined endmembers. Then the projection with the maximum value is selected as the new endmember.

To unmix highly mixed HSI data, the minimum volume constrained non-negative matrix factorization (MVC-NMF) [109] extracts the endmembers by minimizing both the reconstruction error and the simplex volume that circumscribing the data scatter space determined by the estimated endmembers. The minimization of the reconstruction error serves as an external force to enlarge the simplex volume until it contains all the data points, while the minimization of simplex volume serves as an internal force to reduce the size of the simplex volume as compact as possible. The variable splitting augmented Lagrangian (SISAL) approach [14] applies soft constraints to enforce the spectral vectors to belong to the convex hull of the endmembers signatures. The minimum volume simplex analysis (MVSA) [94] fits a minimum volume simplex to the hyperspectral data by constraining the abundance
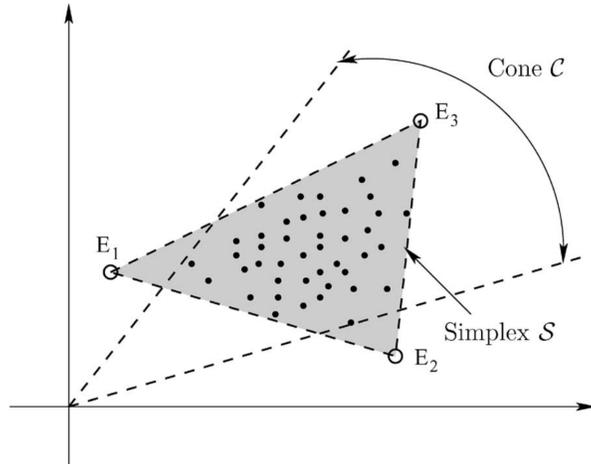
Figure 2.2: Geometric illustration of the simplex that circumscribes the given data denoted by the dots [109]. The endmembers are the vertices of the simplex.

fractions belonging to the probability simplex. The optimization problem is solved by implementing a sequence of quadratically constrained subproblems using the interior point method. Multilayer Nonnegative Matrix Factorization (MLNMF) [122] decomposes the given data iteratively with multiple layers. The sparseness constraint is applied on both the endmembers and abundance matrices in each layer. In this way, the matrix can be sparsely decomposed even the given matrix is not sparse. Robust collaborative non-negative matrix factorization (R-CoNMF) [95] solves the problem by removing the redundant endmember with collaborative regularization.

2.2.3 Statistical based Unmixing

In addition to the geometrical based approaches, statistical based approaches have also drawn much attention recently. According to [17], they are powerful alternatives to unmix highly mixed data although they usually suffer from higher computational cost. Miao *et al.* [110] proposed an unsupervised decomposition method based on the maximum entropy principle. The method incorporates the physical constraints in a natural way and achieves optimal solution by maximizing the entropy of the abundance. Bayesian is one of the most popular frameworks because it can easily impose priors to constrain the solution space. Dobigeon *et al.* [38] estimated the endmembers and abundance by iteratively maximizing the negative log-posterior distribution. In [39], independent Gamma distributions are selected as priors

for the endmembers which naturally handles the non-negativity constraint. In order to naturally incorporate the sum-to-one constraint, Nascimento and Bioucas-Dias [112] modeled the abundance vector as mixtures of Dirichlet densities, which automatically meets the non-negative and sum-to-one constraints. Then the problem is optimized through the cyclic minimization algorithm.

2.2.4 Deep learning based Approaches

Like many other fields of studies, deep learning-based approaches have also thrived in several applications of remote sensing. However, these achievements are limited to supervised learning structures that are mainly based on the convolutional neural network (CNN). For unsupervised spectral unmixing purpose, there has been no breakthrough so far. Due to the unsupervised nature of the unmixing problem, autoencoder, which is an unsupervised learning structure, becomes a natural starting point for investigation. Furthermore, due to the non-negativity requirements of the unmixing problem, the networks that used to address such problem belonging to the category of part-based autoencoder.

Here, we briefly review some recently published approaches that learn such part-based representations through autoencoders. In 2010, Lemme *et al.* [93] developed an online training algorithm with tied encoder and decoder layers. The tied weights are regularized to be non-negative and the hidden layer is encouraged to be sparse. In 2014, Chorowski *et al.* [34] further showed that, the interpolatable power of a network is improved by constraining neurons' weights to be non-negative. Furthermore, combining a softmax on the output layer of the non-negative network can extract representations with understandable physical meanings. In 2016, a non-negative constrained autoencoder was proposed by Hosseini-Asl *et al.* [67] to learn part-based representations by introducing the sparsity to the hidden layers and non-negativity to both the encoder and decoder layers. Although these designs learn the optimal features that achieved promising results for classification, there is no sum-to-one constraint imposed in any of the networks. It makes the representations not truly part-based, and leads to potential instability during training.

Guo *et al.* [58] proposed an autoencoder cascade to solve the unmxing problem, which concatenates a marginalized denoising autoencoder with a non-negative sparse autoencoder.

However, the accuracy of the estimated endmembers is limited by the usage of the tied-weights structure. Also, the reconstruction error is high due to the error introduced in the cascading layers, i.e., the input to the unmixing network is the output of the denoiser, which introduced additional error. Furthermore, the network is easy to be deactivated due to its element-wised way to optimize the network. Ozkan *et. al.* [119] proposed a sparse autoencoder for nonlinear unmixing, which replaces inner product operators with spectral angular mapper to obtain more discriminative representations and enforce the angular similarity between the reconstruction data and the input data. However, both networks use a tied-weights structure. We will show in Sec. 3.3 that due to the constraints of the tied-weights, the representations learned by existing designs carry limited merits.

2.3   Hyperspectral Image Super-Resolution

2.3.1   Multispectral Pansharpening

Traditionally widely utilized multispectral pan-sharpening methods can be roughly categorized into two groups, the component substitution (CS) and the multi-resolution analysis (MRA) based approaches. Generally, CS–based approaches [145] project the given data onto a predefined space, where the spectral information and spatial information are separated. Subsequently, the spatial component is substituted with the one extracted from PAN [25, 3]. MRA based approaches achieve the spatial details by first applying a spatial filter to the HR images. Then the spatial details are injected into the LR HSI [104, 131, 19, 2, 100]. Although these traditional pan-sharpening approaches can be extended to solve the HSI-SR problem, they usually suffer from severe spectral distortions [100, 4, 37].

2.3.2   Bayesian based Pansharpening

Bayesian approaches estimate the posterior distribution of the HR HSI given LR HSI and HR MSI. The unique framework of Bayesian offers a convenient way to regularize the solution space of HR HSI by employing a proper prior distribution. Different methods vary according to the different prior distributions adopted. Wei *et al.* proposed a Bayesian

Naive method [157] based on the assumption that the representation coefficients of HR HSI follow a Gaussian distribution. However, this assumption does not always hold especially when the ground truth HR HSI contains complex textures. Instead of using Gaussian prior, dictionary based approaches solve the problem under the assumption that HR HSI is a linear combination of properly chosen over-complete dictionary and sparse coefficients [156]. Simoes *et al.* proposed HySure [135], which takes into account both the spatial and spectral characteristics of the given data. This approach solves the problem through vector based total variation regularization. Akhtar *et al.* [4] introduced a non-parametric Bayesian strategy to solve the HSI-SR problem. The method first learns a spectral dictionary from LR HSI under the Bayesian framework. Then it estimates the spatial coefficients of the HR MSI by Bayesian sparse coding. Eventually, the HR HSI is generated by combining the spatial dictionary with the spatial coefficients.

2.3.3   Matrix Factorization based Pansharpening

Matrix factorization based approaches have been actively studied recently [77, 166, 43, 87, 150], with Kawakami *et al.* [77] being the first that introduced matrix factorization to solve the HSI-SR problem. The method learns a spectral basis from LR HSI and then use this basis to extract sparse coefficients from HR MSI with non-negative constraints. Similar to Bayesian based approaches, the HR HSI is generated by linearly combining the estimated bases with the coefficients. Yokoya *et al.* [166] decomposed both the LR HSI and HR MSI to achieve the optimal bases and coefficients alternatively. In their approach, both the bases and coefficients are also enforced to be non-negative. Wycoff *et al.* [161] solved the problem with alternating direction method of multipliers (ADMM). Lanaras *et al.* [87] further improved the fusion results by introducing a sparse constraint. However, most methods [166, 161, 87] are based on the same assumption that the down sampling function between the spatial coefficients of HR HSI and LR HSI are known beforehand. This assumption is not always true.

2.3.4 Deep learning based Super-Resolution

Deep learning attracts an increasing attention for natural image super-resolution since 2014, when Dong *et al.* first introduced convolution neural network (CNN) to solve this problem and demonstrated state-of-the-art restoration quality [41]. Ledig *et al.* proposed a method based on generative adversarial network and skipped residual network [62]. The method employed perceptual loss through VGG network [136, 75] which is able to recover photo-realistic textures from heavily down-sampled images [91]. Natural image SR methods usually works on up to 4× upscaling.

There have been three attempts to address the MSI pan-sharpening (MSI-PAN) problem with deep learning in a supervised fashion. In 2015, a modified sparse tied-weights denoising autoencoder was proposed by Huang *et al.* [70] to enhance the resolution of MSI. The method assumes that the mapping function between LR and HR PAN are the same as the one between LR and HR MSI. Masi *et al.* proposed a supervised three-layer SRCNN [107] to learn the mapping function between LR MSI and HR MSI. Similar to [107], Wei *et al.* [158] learned the mapping function with deep residual network [62].

These deep learning based methods, including natural image SR and MSI-PAN are all supervised, their adoption on HSI-SR remains a challenge due to two reasons. First, they are designed to find a end to end mapping function between the LR images and HR images under the assumption that the mapping function is the same for different images. However, the mapping function may not be the same for images acquired with different sensors. Even for the data collected from the same sensor, the mapping function for different spectral bands may not be the same. Thus the assumption may cause severe spectral distortion. Second, training a mapping function is a supervised problem which requires a large dataset, the down-sampling function, and the availability of the HR HSI, which are not realistic for HSI.

2.4 Anomaly Detection

2.4.1 RX based Approaches

Among all the anomaly detection algorithms, the global RX detector proposed by Reed and Yu [123] is one of the most popular statistical approaches. RX assumes that the probability density function of the background can be modeled as a Gaussian distribution, and the mean and covariance of which can be estimated by the entire data set, since anomaly components are sparse, thus would not affect the parameter estimation. With the estimated parameter, the anomalies can be identified by measuring the Mahalanobis distance of a pixel vector to its background. One potential issue of RX is that normally the background components contained in real HSI is nonuniform, which cannot be simply modeled using the Gaussian distribution.

To overcome these limitations, several algorithms have been proposed based on the basic global RX, including the regularized-RX [114], subspace RX (SSRX) [128] and local RX [142]. The regularized-RX [114] regularizes the covariance matrix estimated by HSI pixels, which restricts the possible matched filters to a subset. In SSRX, the global RX is applied on a few number of principal component analysis (PCA) bands. However, similar to the global RX, both regularized-RX and SSRX detectors calculate the background statistics using the entire image cube, which may attenuate the difference between anomalies and their background. Local RX performs the global RX in a fixed local window, which is able to improve the detection rate. But the drawback is that it may struggle with isolated noisy pixels, which would increase the false alarm rate drastically. Also the size of the local window needs to be defined according to the size of anomaly, which is usually unknown.

2.4.2 Representation based Approaches

In addition to the statistical approaches, another branch of anomaly detection is the representation-based method which assumes that hyperspectral image pixels can be represented by a dictionary [30, 28]. A joint sparsity model was proposed in [29] where the pixels within a small neighborhood can be simultaneously represented by a linear combination of

pre-trained dictionary of both target and background classes. Then the sparse coefficient vectors are used to detect anomalies and the background. Similarly, Li and Du proposed a collaborative representation detection method (CRD) [96], which states that if a pixel belongs to the background, it can be approximately represented by a linear combination of its spatial neighborhood. A distance-weighted regularization matrix is introduced in the optimization process to adjust the weight of each pixel.

2.4.3 Low-rank based Approaches

Li *et al.* [97] proposed a method based on robust principal component analysis (RPCA) [98, 20], Tucker decomposition and unmixing [109]. RPCA was introduced as the first step to extract a dense low-rank tensor. However, RPCA is based on the assumption that data have a single subspace. Due to the existence of mixed pixels in HSI data, HSI pixels are usually a linear combination of multiple endmembers. Therefore, RPCA may not be appropriate for this scenario. Wang *et al.* [154] proposed a method based on unmixing [109] and RPCA [98], which has been performing well in detecting anomalies, but often with high false alarm rate. This is because RPCA performs effectively only for isolated pixels. And if the noise is isolated, it would have similar property as anomaly, thus is easy to be detected as anomaly.

To reduce the influence of noise, Zhang *et al.* [173] proposed a method based on Go Decomposition (GoDec) [180]. This algorithm decomposes the HSI data into three structures, i.e., background, anomaly, and noise, instead of just background and anomaly. The noise in HSI is estimated first during the GoDec decomposition [180]. Then the anomaly is detected by applying the Global RX on the sparse matrix. Unfortunately, anomalies are prone to be identified as noise and thus detection rate is reduced in practice. Low-Rank Representation (LRR) [99] is another way to decompose the data. It uses the input data as the dictionary to separate the background and the anomaly. During the procedure, the l_{21} constraint is applied on the anomaly matrix to solve the problem when data lie in multiple subspaces.

Chapter 3

Spectral Unmixing

3.1 Introduction

One of the challenging issues in HSI analysis is the handling of the so-called "mixed pixel", where each pixel tends to cover more than one constituent material within the instantaneous field of view of the sensor. To improve performance, subpixel-level accuracy is usually desired, making the process of *spectral unmixing* essential in HSI analysis. Spectral unmixing refers to the process of deriving the constituent components (i.e., *endmembers*) and their fractional proportions (i.e., *abundance*) from the mixed pixels [17, 79]. It is well known to be a difficult and highly ill-posed inverse problem and the estimation is highly sensitive to noise.

In general, spectral unmixing involves three major components: the estimation of the number of endmembers, the extraction of endmembers, and the estimation of abundance. In this section, we follow majority of the previous works and assume a linear mixing model where different endmembers do not interfere with each other [17, 79, 6]. Under this model, the observed spectrum at a single pixel is formulated as a linear combination of individual endmember spectra, with the mixing coefficients being the *abundance*. Based on this formulation and the physical meaning it implies, both the abundance and endmember spectra need to be non-negative and the mixing coefficients need to be sum to one. The non-negativity and sum-to-one constraints are generally used in solving the linear unmixing problem to regularize the solution space.
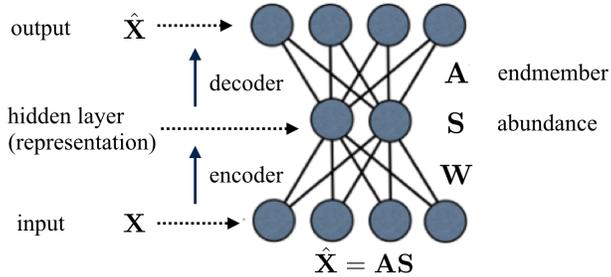
Figure 3.1: The correspondence between the autoencoder model and the spectral unmixing procedure.

In the past few decades, there have been numerous approaches developed to solve the linear unmixing problem, including both semi-supervised and unsupervised approaches. Semi-supervised approaches solve the problem by sparse regression [124, 71]. Given the set of endmembers known in advance, it estimates the optimal subset of endmember collections that best fits the input data. However, in most scenarios, the endmembers in HSI are unknown, thus the unmixing needs to be performed in an unsupervised fashion. For most of these traditional unsupervised approaches, noise remains a challenge which could dramatically decrease the unmixing accuracy, especially when the number of endmembers is incorrectly estimated.

In this section, we intend to fill in this gap by establishing a connection between unsupervised deep learning and unsupervised spectral unmixing, where the unmixing procedure can be explained as finding a set of low-dimensional representations (i.e., abundance) that reconstruct the data with their corresponding bases (i.e., endmembers). This interpretation facilitates the application of autoencoder-based unsupervised deep learning models in solving spectral unmixing problems in an unsupervised fashion. An autoencoder learns the low-dimensional representation of the data automatically by minimizing the reconstruction error [10, 55]. As shown in Fig. 3.1, the autoencoder usually consists of an encoder which extracts the representations of the data, and a decoder which reconstructs the data from representations. We observe that the decoder process resembles the unmixing procedure, which makes it a good model to solve the unmixing problem (See more details in Sec. 3.2). However, due to the non-negativity and sum-to-one requirements, to solve the unmixing problem with autoencoder, the weights and representation layer of the network have to

be limited to non-negative, and the representation layer needs to satisfy the sum-to-one constraint. With the non-negative constraints, such autoencoder network is often referred to as the *part-based autoencoder*, and the representations measured by the network as *part-based representations.*

Recently, there have been several attempts to learning such part-based representations through part-based autoencoders [93]. However, part-based autoencoders have not performed well for spectral unmixing. This is mainly because most existing part-based autoencoders adopt a tied-weights structure, i.e., the encoder weight ($\mathbf{W}$) equals to the transpose of the decoder weight ($\mathbf{A}$) as shown in Fig. 3.1. In Sec. 3.3, we will elaborate on why this is an inappropriate assumption that would ultimately lead to ineffective unmixing result. In addition, none of the existing works have an effective mechanism to incorporate the sum-to-one and non-negative constraints. For example, the non-negativity constraint is usually enforced by de-activating the weights that are negative. This would have the side effect of removing the connection permanently during the training procedure. On top of these challenges, noise remains an issue for the network to learn meaningful part-based representations.

To overcome the drawbacks of existing approaches, in this section, I propose an untied denoising autoencoder with sparsity to solve the problem of spectral unmixing, refereed as uDAS.

3.2 Problem Formulation

According to the linear mixing model [17], the observed spectral reflectance $\mathbf{X}$ can be formulated as

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N} \tag{3.1}$$

where $\mathbf{X} \in \mathbb{R}^{l \times n}$ is the observed mixture with each column representing the reflectance reading of one pixel (or data samples) in l spectral bands; $\mathbf{N} \in \mathbb{R}^{l \times n}$ represents the noise; $\mathbf{A} \in \mathbb{R}^{l \times c}$ is the endmember matrix with each column representing one of the c endmember signatures of, again, l spectral lands; and $\mathbf{S} \in \mathbb{R}^{c \times n}$ denotes the corresponding abundance matrix, with each column representing the mixing coefficient of the c endmembers in making
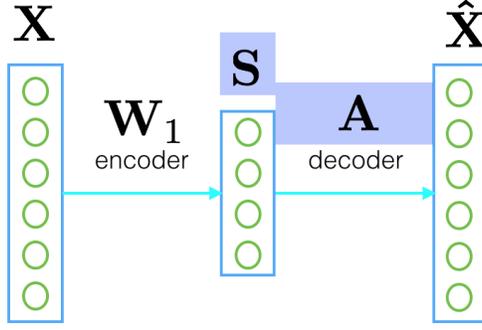
Figure 3.2: Formulation of the unmixing problem with autoencoder.

one mixed pixel. To solve the spectral unmixing problem, three physical constraints need to be satisfied, i.e., the endmember $\mathbf{A} \geq \mathbf{0}$, and the abundance vectors $\mathbf{s}_i \in \mathbb{R}^c$ meet the non-negativity requirement $(\mathbf{s}_i \geq \mathbf{0})$ and the sum-to-one requirement $(\sum_{i=1}^c s_{ij} = 1)$.

To solve the unmixing problem, a three-layer autoencoder network is constructed, as shown in Fig. 3.2. Traditionally, autoencoder has been used for dimensional reduction purpose, where the hidden layer output is the learned low-dimensional representation of the input data that are used to reconstruct the data. Similarly, the unmixing process can also be explained as finding low-dimensional representation, i.e., abundance $\mathbf{S}$ that reconstructs the data. Therefore, theoretically we should be able to perform unmixing using autoencoder.

The network consists of two parts:

- an encoder $f(x)$ that encodes the input data, $\mathbf{X}$, to representations, $\mathbf{S}$, (i.e., hidden layer output or abundance) as shown in Eq. 3.2,

$$\mathbf{S} = f(\mathbf{X}) = \sigma(\mathbf{W}_1 \mathbf{X}) \tag{3.2}$$

  where $\sigma(x)$ is the component-wise activation function and $\mathbf{W}_1$ represents the encoder weight that connects the input layer and the hidden layer,

- a decoder $g(\mathbf{S})$, which reconstructs the data using $\mathbf{S}$, as shown in Eq. 3.3,

$$\hat{\mathbf{X}} = g(\mathbf{S}) = \mathbf{AS} \tag{3.3}$$

where $\mathbf{A}$ is the decoder weight matrix (endmember) representing the weights connecting the hidden layer and the output layer, and $\hat{\mathbf{X}}$ denotes the reconstructed data.

Note that the decoder $\mathbf{A}$ and the hidden layer $\mathbf{S}$ naturally match with the endmember and abundance, respectively, in the unmixing problem. The network learns the weights and representations that reconstruct the data by minimizing the average reconstruction error:

$$J(\mathbf{W}_1, \mathbf{A}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|g(f(\mathbf{x}_i)) - \hat{\mathbf{x}}_i\|^2 \qquad (3.4)$$

Note that we do not use bias in our network design because bias tends to be a large negative value during the training procedure. Given this matching and formulation, we need to further investigate how to incorporate the non-negative and sum-to-one constraints in order to solve the unmixing problem.

For simplicity, the notations employed in this section are tabulated in Table 3.1.

3.3 Potential Issues in Existing Part-based Autoencoder Design for Spectral Unmixing

Given the basic autoencoder model, there are two key issues we need to investigate in order to design an efficient and effective deep learning solution to the unmixing problem. The first is about the network structure; all existing part-based autoencoder designs use a tied-weight assumption as well as a non-negative encoder/decoder weight; are these assumptions appropriate? The second is tailored toward the spectral unmixing problem where two physical constraints need to be satisfied, e.g., sum-to-one and non-negativity; how to incorporate these constraints in the network?

In this section, we present some important findings on the network design. Using the method of proof by counterexample, we show that in order to learn the endmember $\mathbf{A}$ and abundance $\mathbf{S}$ correctly, 1) the encoder of the network cannot be tied with the decoder, and 2) the encoder should not be constrained to be non-negative.
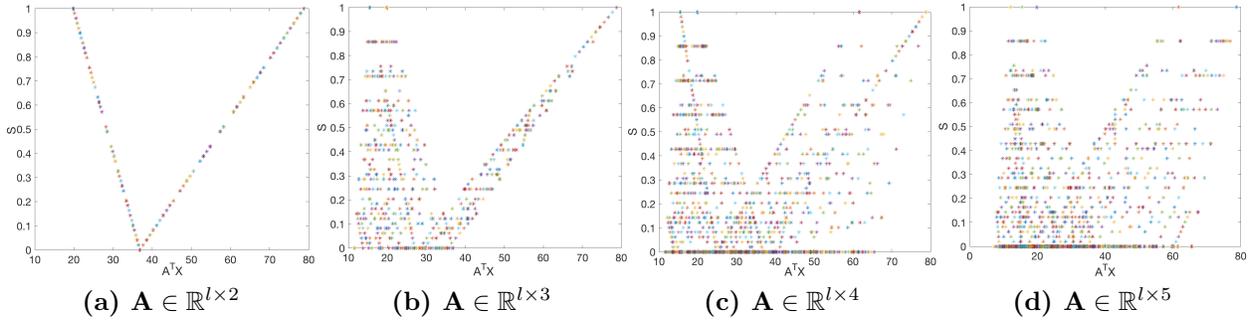
Figure 3.3: The mapping function σ that forces $\mathbf{S} = \sigma(\mathbf{A}^T\mathbf{X})$ when the number of endmembers c is 2, 3, 4 and 5.

Given a list of randomly selected $\mathbf{A} \in \mathbb{R}^{l \times c}$, we generate a synthetic non-negative observation dataset $\mathbf{X} \in \mathbb{R}^{l \times n}$ of n samples by linearly combining the given endmember $\mathbf{A}$ using a set of randomly generated abundance $\mathbf{S}$ but satisfying the sum-to-one and non-negativity constraints. That is, $\mathbf{X} = \mathbf{AS}$. Suppose we design a three-layer autoencoder for unmixing purpose, where $\mathbf{W}_1 > 0$ is the weight for the encoder network and $\mathbf{A} > 0$ is the weight for the decoder network. In addition, $\mathbf{S} \geq 0$ and each column of $\mathbf{S}$ sums to 1.

3.3.1 The Appropriateness of the Tied-Weights Assumption

The tied-weight assumption assumes the encoder $\mathbf{W}_1$ and the decoder $\mathbf{A}$ of the network are tied, i.e., $\mathbf{W}_1 = \mathbf{A}^T$. Then we have

$$\mathbf{S} = \sigma(\mathbf{W}_1\mathbf{X}) = \sigma(\mathbf{A}^T\mathbf{X}) \tag{3.5}$$

where σ is the activation function like sigmoid. Since we already know $\mathbf{S}$ and $\mathbf{A}$, the activation function that maps $\mathbf{A}^T\mathbf{X}$ to $\mathbf{S}$ can be drawn in Fig. 3.3. Note that the y-axis denotes the values of each element in $\mathbf{S}$ and the x-axis denotes the values of each element in $\mathbf{A}^T\mathbf{X}$. In order to better observe the behavior of the activation function, we illustrate the cases where the number of endmembers is 2, 3, 4 and 5. In all cases, we observe a peculiar non-monotonic relationship between $\mathbf{S}$ and $\mathbf{A}^T\mathbf{X}$, which is very much against what an activation function should behave, i.e., monotonic increasing or decreasing. Note that there does exist activation

(a)   **(b)**   **(c)**   **(d)**
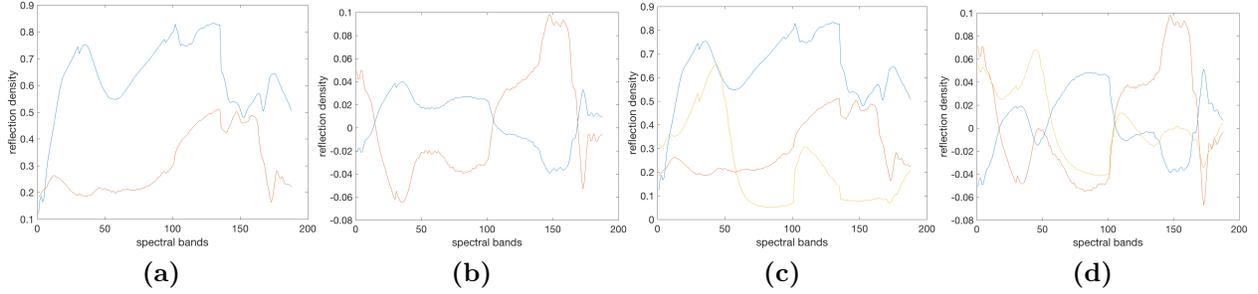
Figure 3.4: The inappropriateness of assuming both $\mathbf{W}_1$ and $\mathbf{A}$ to be non-negative. Given two examples of $\mathbf{A}$ in (a) Decoder $\mathbf{A} \in \mathbb{R}^{188 \times 2}$ and (c) Decoder $\mathbf{A} \in \mathbb{R}^{188 \times 3}$, the corresponding $\mathbf{W}_1$ is calculated as the pseudo inverse of $\mathbf{A}$, i.e., $\mathbf{W}_1 = \mathbf{A}^+$, as shown in (b) and (d) respectively. We observe from (b) and (d) that the elements of $\mathbf{W}_1$ obtained from $\mathbf{W}_1 = \mathbf{A}^+$ are not all non-negative, showing the assumption that both encoder and decoder weights need to be non-negative is inappropriate.

functions, such as max out [56], where non-monotonic increment can be modeled. However, these functions are computationally-expensive to calculate.

To conclude, since under the tied-weights assumption, the activation function will not be monotonically increasing; the assumption must be inappropriateness.

3.3.2   The Appropriateness of the Non-Negative Weight Assumption

If we remove the tied-weights assumption, the network should satisfy $\mathbf{X} = \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X})$ where $\mathbf{W}_1^T \neq \mathbf{A}$. In this section, we further investigate the appropriateness of assuming both the encoder and decoder weights to be non-negative. Without loss of generality, let's assume the activation function to be the simplest identity function, i.e., $\sigma(x) = x$. Then, we have $\mathbf{X} = \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X}) = \mathbf{A}\mathbf{W}_1\mathbf{X}$. In order for the equality relationship to hold, $\mathbf{A}\mathbf{W}_1 = \mathbf{I}$. That is, given $\mathbf{A}$, the correct $\mathbf{W}_1$ should be the pseudo inverse of $\mathbf{A}$, i.e., $\mathbf{W}_1 = \mathbf{A}^+$. This relationship is depicted in Fig. 3.4, where we observe from Figs. 3.4b and 3.4d that although all elements of $\mathbf{A}$ are set to be non-negative (due to physical constraints of endmembers), not all the elements of $\mathbf{W}_1$ are non-negative. Therefore, by enforcing $\mathbf{W}_1 \geq \mathbf{0}$, the network would fail to learn the correct $\mathbf{W}_1$, leading to inaccurate estimation of endmembers and abundances.

Based on the analyses above, we summarize our important findings as follows: In order to obtain more accurate estimation of endmember and abundance,

- the network should not have tied-weights encoder and decoder, i.e., $\mathbf{W}_1 \neq \mathbf{A}^T$, and

- although the decoder weight, $\mathbf{A}$, should be non-negative due to the physical constraints of endmembers, the encoder weight, $\mathbf{W}_1$, however, does not possess this property.

In our design, we use two different weight matrices, $\mathbf{W}_1$, the weights connecting the input layer $\mathbf{X}$ to the hidden layer $\mathbf{S}$, and $\mathbf{A}$, the weights connecting the hidden layer and the output layer $\hat{\mathbf{X}}$. The detailed design is elaborated in the next section.

3.4 uDAS: untied Denoising AutoEncoder with Sparsity

The network model of the proposed uDAS is illustrated in Fig. 3.5. It possesses the following three unique characteristics: First, the encoder and the decoder of the network are independent, and only the decoder is required to be non-negative. This has been analyzed in Sec. 3.3. Second, instead of cascading a denoising layer as a preprocessing step, we propose a denoising constraint $\mathbf{W}_n$ on the decoder $\mathbf{A}$ of the network to reduce reconstruction error. This will be discussed in Sec. 3.4.1. Third, since the accurate estimation of the number of endmembers plays an important role in unmixing, especially when the input data is noisy, we introduce an l_{21} norm on the encoder, which is equivalent to reducing the redundant rows in $\mathbf{W}_1$ and the hidden layer representation (abundance) $\mathbf{S}$. This effectively reduces the redundant endmembers. See Sec. 3.4.2 for detail.

3.4.1 Denoising Constraint on the Decoder

Recent studies [79, 17, 95] have shown that the presence of noise will degrade the performance of unmixing extensively. Existing unmixing approaches all have a denoising component as a pre-processing step during the unmixing practice. However, a concatenated denoising component will add additional reconstruction error on top of the reconstruction error from
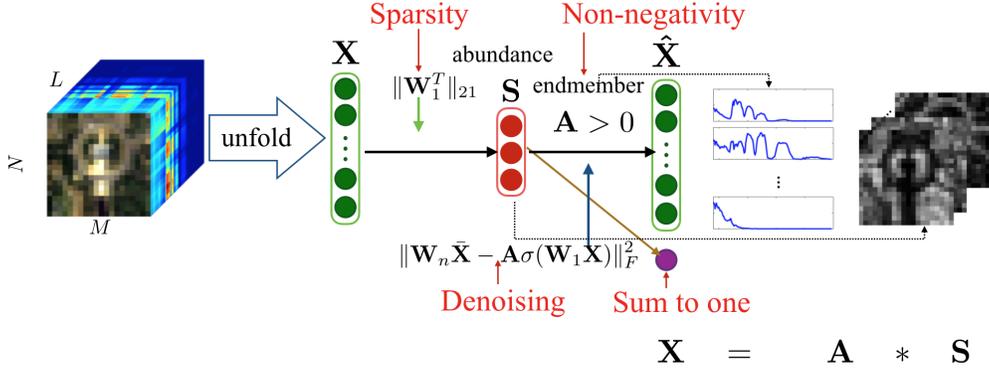
Figure 3.5: The network structure of the proposed uDAS.

the autoencoder. Therefore, we design an end-to-end approach where denoising is integrated in the objective function of the autoencoder design and can be treated as a denoising constraint. In the following, we first briefly review the marginalized denoising autoencoder (mDA) on which the designed denoising constraint is based.

Marginalized Denoising Autoencoder

The proposed denoising constraint is based on the marginalized denoising autoencoder (mDA), which is capable of marginalizing out the corruption during training. The mDA assumes that there exists a linear transformation function $\mathbf{W}_n = [\mathbf{W}, \mathbf{b}]$, which can reconstruct the clean data by reducing the squared loss between the original clean data $\hat{\mathbf{x}}_i$ and the measured noisy data, $\mathbf{x}_{i,j}$ as shown in Eq. (3.6):

$$L(\mathbf{W}_n) = \frac{1}{2\bar{m}n} \sum_{j=1}^{m} \sum_{i=1}^{n} \|\hat{\mathbf{x}}_i - \mathbf{W}_n \mathbf{x}_{i,j}\|^2, \tag{3.6}$$

where n is the number of pixels in the image, $\bar{m}$ is the number of noisy measurements randomly generated, and the vector $\mathbf{x}_{i,j}$ denotes the jth corrupted version of the original input $\hat{\mathbf{x}}_i$. The core idea of mDA is by randomly generating $\bar{m}$ noisy measurements and stacking them up, then the clean data can be reconstructed by marginalizing these noisy measurements. Let $\mathbf{X}$ denote the corrupted data cube (i.e., the noisy input), $\hat{\mathbf{X}}$ denote the clean data cube, $\tilde{\mathbf{X}}$ denote the stacked $\bar{m}$ different corrupted $\mathbf{X}$, and $\mathbf{X}_r$ denote a stack of
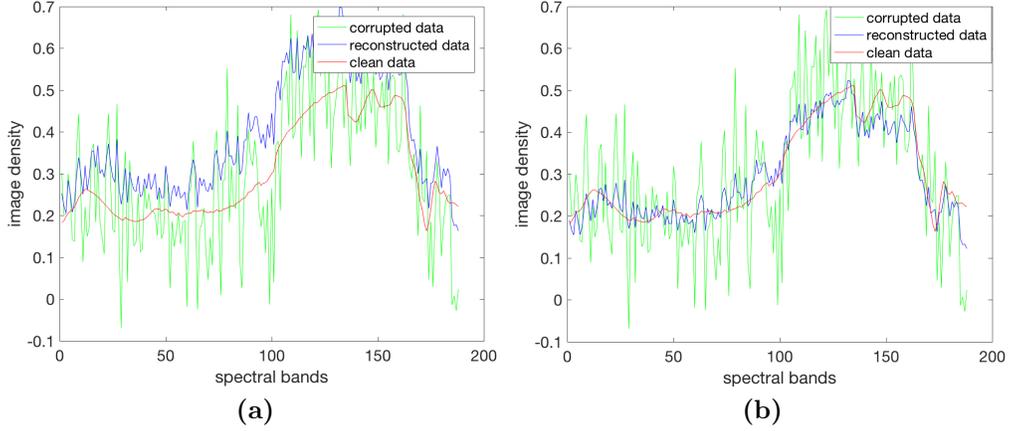
Figure 3.6: Reconstructed results from noisy data with SNR = 10dB using mDA. (a) Result from the original mDA. (b) Result from the scaled mDA.

$\hat{\mathbf{X}}$ duplicated for $\bar{m}$ times. The objective function in Eq. (3.6) can be converted to

$$L(\mathbf{W}_n) = \frac{1}{2\bar{m}n}\mathrm{tr}[(\mathbf{X}_r - \mathbf{W}_n\tilde{\mathbf{X}})^T(\mathbf{X}_r - \mathbf{W}_n\tilde{\mathbf{X}})]. \tag{3.7}$$

When $\bar{m} \to \infty$, the corruption is marginalized and the transformation function $\mathbf{W}_n$ can be solved using a closed–form solution only with respect to the input noisy data. More mathematic details of the algorithm can be found in [26].

One issue introduced by mDA is that the solution is related to the corruption probability p. When p is large, the resulted reconstructed data tend to be scaled up. Fig. 3.6a shows one pixel of the reconstructed HSI using the original mDA. We can observe that the reconstructed data is scaled up to a large extend. To fix this problem, we reduce the reconstructed signal with $1 - p$, as shown in Eq. (3.8). With this procedure, the reconstructed data is closer to the clean data, as shown in Fig. 3.6b. We refer to this modification as scaled mDA in later sections.

$$\hat{\mathbf{X}} = (1 - p)\mathbf{W}_n\mathbf{X} \tag{3.8}$$

Proposed Denoising Constraint

If we perform the unmixing directly on the data acquired from scaled mDA, the reconstruction error will be large because there would be two reconstruction errors, i.e., the error from the scaled mDA and the error from the unmixing autoencoder. To better handle noise and reduce the reconstruction error simultaneously, instead of cascading the scaled mDA on top of the proposed autoencoder structure, we integrate the denoising feature in our network by introducing a denoising constraint $\mathbf{W}_n$.

Given noise-corrupted input data $\mathbf{X}$, the reconstructed clean data can either be defined as $\hat{\mathbf{X}} = \mathbf{W}_n\mathbf{X}$ using scaled mDA or $\hat{\mathbf{X}} = \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X})$ through the autoencoder structure. Therefore, the relationship between the transformation function $\mathbf{W}_n$ and the proposed network can be defined as:

$$\mathbf{W}_n\bar{\mathbf{X}} = \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X}) \tag{3.9}$$

where $\bar{\mathbf{X}}$ is the augmented data defined in Eq. (3.10) with $\delta = 1$,

$$\bar{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}} \\ \delta\mathbf{1}_n^T \end{bmatrix} \quad \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ \delta\mathbf{1}_c^T \end{bmatrix} \tag{3.10}$$

and $\mathbf{W}_n = (1-p)[\mathbf{W}, \mathbf{b}]$ is the transformation matrix in scaled mDA including the bias $\mathbf{b}$.

Combined with the denoising constraint, the objective function for $\mathbf{A}$ can be written as:

$$J(\mathbf{A}) = \frac{1}{2}\|\mathbf{A}\sigma(\mathbf{W}_1\mathbf{X}) - \mathbf{X}\|_F^2 + \lambda\|\mathbf{W}_n\bar{\mathbf{X}} - \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X})\|_F^2. \tag{3.11}$$

3.4.2 l_{21} Constraint on the Encoder

According to previous research [12, 71, 138, 47, 7], estimating the number of endmembers is another crucial step in the spectral unmixing problem besides denoising.

When the data is clean, the proposed network model can easily estimate the correct number of endmembers, i.e., the number of rows in $\mathbf{W}_1$ which is also the number of hidden nodes of the network. Figure 3.7 shows the experimental results of the estimated $\mathbf{A}$ and its corresponding $\mathbf{W}_1$ on clean data. We can observe that the redundant rows of $\mathbf{W}_1$ (i.e., the
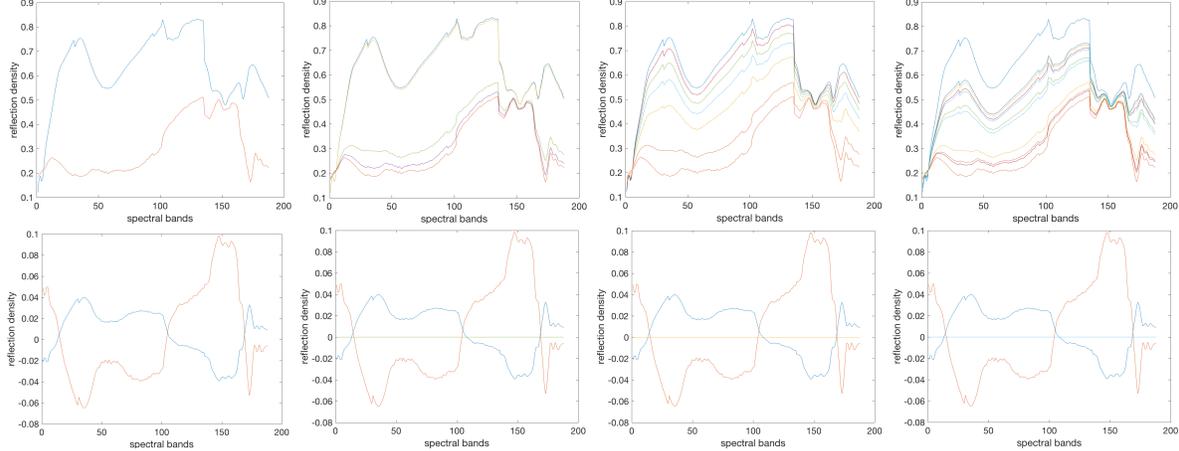
Figure 3.7: Estimated $\mathbf{A}$ and the corresponding $\mathbf{W}_1$ given clean data with $c = 2$. Left four figures are the estimated decoder $\mathbf{A}$ given the number of endmembers $c = 2, 5, 10$ and 20, respectively. The right four figures show the corresponding encoder $\mathbf{W}_1$ of the network.

extra number of endmembers we estimated) are almost zero. Thus, the redundant hidden nodes can be removed (or de-activated) easily. However, when the data is corrupted with noise, it becomes more difficult to estimate the correct number of endmembers, as illustrated in Fig. 3.8, where the redundant rows in $\mathbf{W}_1$ are no longer close to zero. Given noisy data and wrong estimated number of endmembers, the performance of unmixing drops drastically.

Based on the experiments, we also observe that, to achieve a better unmixing result, the encoder weight matrix $\mathbf{W}_1$ should be small. For example, in Fig. 3.7, the endmembers are within the range of $[0, 1]$, but the values of the corresponding $\mathbf{W}_1$ are between -0.08 and 0.1.

To better reflect the desired behavior of the encoder weight, that is, $\mathbf{W}_1$ be small and the corresponding rows of $\mathbf{W}_1$ be close to zero for redundant endmembers, we introduce the l_{21} sparsity constraint [117] to improve the unmixing performance.

In this way, when the redundant rows of $\mathbf{W}_1$ are enforced to be zero, the redundant hidden nodes are de-activated, and the redundant rows in the endmember $\mathbf{A}$ are dropped as well. Note that, since our purpose is to remove redundant rows, the l_{21} norm should be applied to $\mathbf{W}_1^T$ instead of $\mathbf{W}_1$. The l_{21} norm also serves to fix the gradient increasing issue caused by the ReLu activation function to be mentioned in Sec. 3.5.

Another important constraint for the unmixing problem is that the abundance vector $\mathbf{s}_i$ should sum–to–one. To meet this condition, in the $\mathbf{W}_1$ learning procedure, we augment the
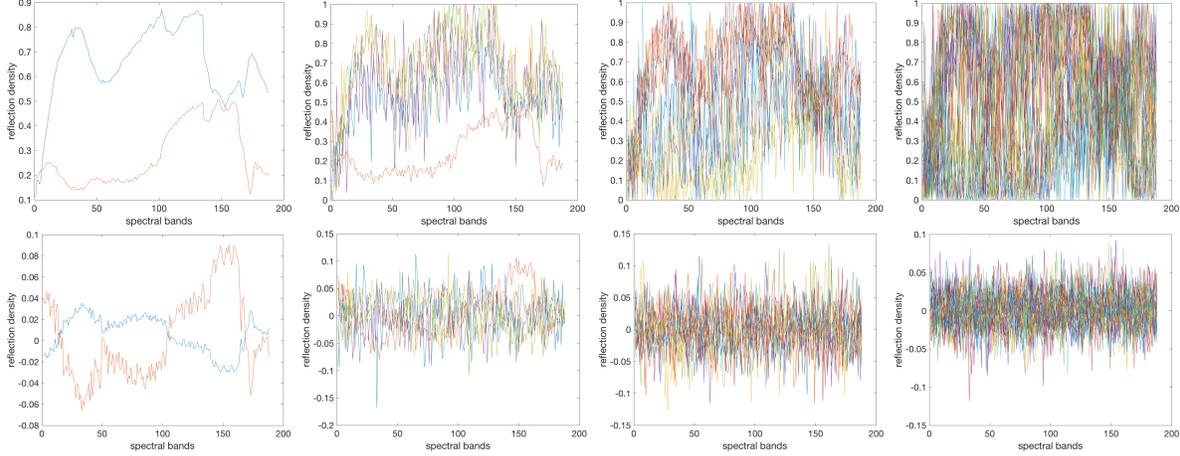
Figure 3.8: Estimated $\mathbf{A}$ and the corresponding $\mathbf{W}_1$ given noisy data with $c = 2$ and SNR = 20 dB. Left four figures are the estimated $\mathbf{A}$ given the number of endmembers $c = 2, 5, 10$ and 20, respectively. The right four figures are the corresponding estimated $\mathbf{W}_1$.

denoised data $\hat{\mathbf{X}}$ and the decoder $\mathbf{A}$ with a constant vector. The augmented data is denoted by $\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$, as defined in Eq. (3.10) with $\delta = 20$ in order to weight the augmented matrix $\mathbf{1}_c^T \mathbf{s}_j = \mathbf{1}_n^T$ more. In this way, the columns of the abundance $\mathbf{s}_j$ are almost sum to one.

To summarize the analyses above, the objective function for $\mathbf{W}_1$ is defined as

$$J(\mathbf{W})_{\mathbf{1}} = \frac{1}{2}\|\bar{\mathbf{A}}\sigma(\mathbf{W}_1\mathbf{X}) - \bar{\mathbf{X}}\|_F^2 + \gamma\|\mathbf{W}_1^T\|_{21} \tag{3.12}$$

Note that the augmentation in Eq. (3.11) is done because there exists bias in the denoising constraint $\mathbf{W}_n$, while the augmentation in Eq. (3.12) is for the purpose of satisfying the sum-to-one constraint.

3.5 Optimization and Implementation Details

For unmixing purpose, the decoding function needs to be a linear mapping function, i.e., $\hat{\mathbf{X}} = \mathbf{AS}$. Thus to achieve the optimal solution, the encoding function should also be linear, as proved by previous research [9]. In addition, the activation function needs to guarantee the hidden layer (abundance $\mathbf{S}$) to be non-negative. Considering these two constraints, we choose the rectified linear unit (ReLu) as the activation function, defined by $\sigma(x) = \max(x, 0)$. When $\mathbf{S} \geq \mathbf{0}$, the encoder is equivalent to a linear mapping function. However, ReLu has

its drawback in terms of increasing the gradient extensively [57]. This issue can be fixed by applying the l_1 or l_2 norm on the weights as proved by recent studies [57]. In our network design, we use the l_{21} norm as discussed in Sec. 3.4.2, which also fixes the problem that caused by ReLu.

To solve the unmixing problem using the proposed method, $\mathbf{A}$ and $\mathbf{W}_1$ are estimated alternatively. Due to the requirements of unmixing, the decoder weight needs to be non-negative, i.e., $\mathbf{A} \geq \mathbf{0}$. problem. Instead of regularizing the weight $\mathbf{A}$, we propose a method that incorporates the Armijo rule [11] which guarantees the non-negativity of $\mathbf{A}$ during the optimization.

$\mathbf{A}$ can be estimated according to the objective function defined in Eq. (3.11) by projecting $\mathbf{A}$ onto the subspace $\mathbf{A} \geq \mathbf{0}$ using

$$\mathbf{A} = \max(\mathbf{A} - \alpha \nabla \mathbf{A}, \mathbf{0}) \tag{3.13}$$

where

$$\begin{aligned} \nabla \mathbf{A} &= (\mathbf{AS} - \mathbf{X})\mathbf{S}^T + \lambda(\mathbf{AS} - \mathbf{W}_n \bar{\mathbf{X}})\mathbf{S}^T \\ \mathbf{S} &= \sigma(\mathbf{W}_1 \mathbf{X}). \end{aligned} \tag{3.14}$$

The learning rate α is selected by the Armijo rule [11]. Then $\mathbf{W}_1$ is estimated by back propagating the reconstruction error from the last layer through $\mathbf{A}$ with the sum-to-one constraint shown in Eq. (3.12).

In the proposed uDAS method, we initialize the decoder $\mathbf{A}$ with endmembers estimated by VCA [113, 95]. The encoder is initialized via $\mathbf{W}_1 = \mathbf{SX}^{-1}$, where $\mathbf{S}$ is learned through the fully constrained least squares (FCLS) [117] method. The whole procedure is described in Algorithm 2.

3.6 Experimental Results

In this section, the proposed algorithm is evaluated with hyperspectral image datasets including both synthetic and real data. In the synthetic data experiments, the properties of the proposed algorithm is fully investigated with both visual and quantitative comparisons.

Algorithm 1 Untied Denoising Autoencoder with Sparsity l_{21} for Spectral Unmixing (uDAS)

Input: X
Output: A, S

  Initialization : initialize the encoder **A** through VCA, and the decoder $\mathbf{W}_1 = \mathbf{SX}^{-1}$, with **S** estimated through FCLS.
  1) Estimate $\mathbf{W}_n$ through scaled mDA.
  while not converge **do**
  2) Fix $\mathbf{W}_1$ and update **A** through
    $\nabla \mathbf{A}^k \quad = (\mathbf{A}^k\mathbf{S} - \mathbf{X})\mathbf{S}^T + \lambda(\mathbf{A}^k\mathbf{S} - \mathbf{W}_n\bar{\mathbf{X}})\mathbf{S}^T$
    $\mathbf{A}^{k+1} \quad = \max(\mathbf{A}^k - \alpha\nabla\mathbf{A}^k, \mathbf{0})$
    $\mathbf{S} \qquad = \sigma(\mathbf{W}_1\mathbf{X})$.
  3) Fix **A** and update $\mathbf{W}_1$ through
    $\nabla\mathbf{W}_1^k \quad = \bar{\mathbf{A}}^T(\bar{\mathbf{A}}\mathbf{S} - \bar{\mathbf{X}}) \cdot \sigma^{-1}(\bar{\mathbf{X}}) + \gamma\mathbf{W}_1^{k^T} \cdot \mathbf{M}^{-\frac{1}{2}}$
    $\mathbf{W}_1^{k+1} \quad = \mathbf{W}_1^k - \beta\nabla\mathbf{W}_1^k$
    $\mathbf{m}_{ij} \qquad = \sum_{j=1}^l \mathbf{w}_{ij}^{k\,2}$
  Check the convergence:
    $\|\mathbf{AS} - \mathbf{X}\|_F^2 < \varepsilon$
  end
  return A, S

We further show that the proposed algorithm is capable of solving real applications using the well-studied AVIRIS Cuprite dataset [52].

3.6.1 Dataset Description

The synthetic images of size $32 \times 32 \times 188$ are generated using a linear mixing of five spectral signatures (endmembers) following the rule that used in [109], in order to simulate a hyperspectral scene with endmembers lie in discrete patches. The spectral signatures are selected from the United States Geological Survey (USGS) library. There are totally 224 spectral bands acquired from wavelength 0.4 to 2.5 μm with an interval of 10 nm. Due to water discarded, 36 bands are abandoned [52, 95]. Thus we use the rest 188 bands, i.e., $\mathbf{A} \in \mathbb{R}^{188 \times 5}$, for unmixing purpose. To create linear mixtures, the entire image is divided into 16 8×8 blocks. To start with, each block contains only pure pixels. That is, each pixel is only made up of one endmember. Then a 7×7 low pass filter is applied on the image cube to simulate the degraded image cube with mixed pixels.

In order to evaluate the robustness of the proposed algorithm, the Gaussian white noise is added to the clean data with the signal-to-noise ratio (SNR) varying from 50 dB to 10 dB, to simulate the degradation of hyperspectral data.

The real data used to evaluate the proposed method is the well-studied AVIRIS Cuprite data [52] with resolution $250 \times 191 \times 224$. Figure 3.13 shows the mineral map generated by USGS in 1995. Similar to the synthetic data, due to water absorption and low SNR, bands 1–2, 105–115 and 150–170 and 224–224 bands are discarded [52, 95]. Thus 188 bands are left for analysis.

3.6.2 Experimental Design and Performance Metrics

We evaluate the proposed uDAS from two perspectives. *First*, we perform comprehensive evaluation on the different components of the proposed approach, i.e, the denoising constraint and the l_{21} constraint, including

- unmixing without any denoiser, with the cascaded scaled mDA denoiser, and with the proposed denoising constraint– this comparison would show that the proposed denoising constraint on $\mathbf{A}$ can increase the accuracy of the estimated endmembers: and

- unmixing without and with the l_{21} constraint– with this experiment, we would show that the proposed l_{21} constraint on $\mathbf{W}_1^T$ is able to effectively reduce the redundant rows of $\mathbf{W}_1$ and thus de-activate the redundant endmembers.

In addition, we compare the proposed algorithm with five representative approaches including minimum volume constrained non-negative matrix factorization (MVC-NMF) [109], multilayer non-negative matrix factorization (MLNMF) [122] , minimum volume simplex analysis (MVSA) [94], and robust collaborative non-negative matrix factorization (R-CoNMF) [95], which are geometrical-based approaches; and Bayesian-based unmixing [38], which are statistical-based approaches.

Besides qualitative comparison via visual inspection, for quantitative comparison, unmixing results from the synthetic data are evaluated using five metrics, including the
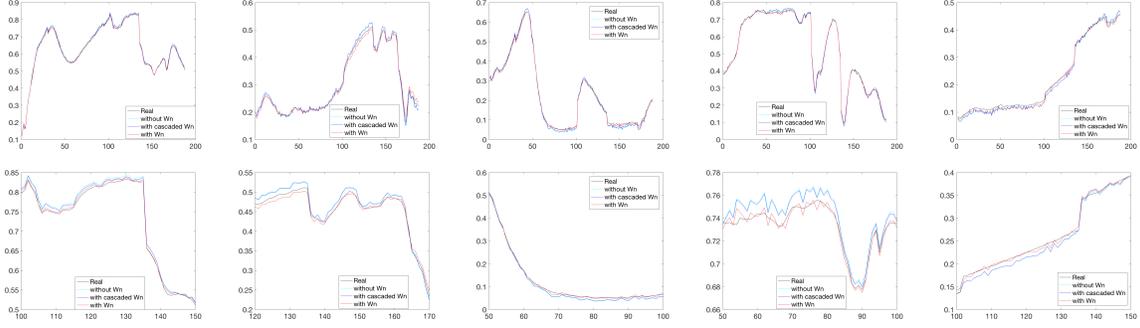
Figure 3.9: Denoising constraint evaluation on synthetic data with SNR = 20 dB. Top: estimated endmembers. Bottom: zoomed in endmembers.

root mean square error (RMSE), abundance angle distance (AAD), abundance information divergence (AID), spectral angle distance (SAD), and spectral information divergence (SID) [109]. The lower the value of the metrics, the better the performance of an algorithm. For the real Cuprite data, the extracted endmembers are compared with the USGS library signatures with SAD in order to show the effectiveness of the proposed algorithm.

3.6.3 Experiments with Synthetic Data

Effects of the Denoising Constraint

In this set of experiments, we would like to evaluate the impact of adding the denoising constraint, as compared to the design using a cascaded denoiser and the design without any denoising mechanisms. Assuming that the number of endmembers $c = 5$ is known, when SNR = 20 dB, the results of the estimated $\mathbf{A}$ are shown in Fig. 3.9. The left column of Fig. 3.9 shows the five extracted endmembers, and the right column of Fig. 3.9 are the zoomed-in version for more clear visualization. We can observe that, with a cascaded scaled mDA denoiser, the extracted endmembers are closer to the ground truth (real) as compared to the results without any denoiser. However, with the proposed denoising constraint, the extracted endmembers are much closer to the ground truth. This is further supported by the following quantitative study.

For quantitative comparison, we calculate the SAD and SID of the estimated endmembers against the ground truth endmembers and list the results in Table 3.2 with two levels of noise added. We observe that, for very noisy data with low SNR, e.g., SNR = 10 dB and SNR = 20

dB, although both the spectral angle and information distance of the extracted endmembers decrease, the proposed denoising constraint still achieves the best result among the three design options. This indicates that denoising is important for spectral unmixing and that using denoising constraint such that only one reconstruction error is introduced instead of two in the cascading case, the proposed algorithm is able to extract more accurate endmembers.

Effects of the Sparsity Constraint

According to the analysis in Sec. 3.4.2, the the number of endmembers would impact the unmixing performance significantly when there exists noise. Thus an l_{21} constraint is introduced to reduce the redundant endmembers, i.e., redundant rows of abundance $\mathbf{S}$ and $\mathbf{W}_1$. In this section, we use synthetic data with SNR $= 20$ dB as an example to evaluate the effectiveness of the l_{21} constraint. In the following experiments, $\gamma = 10^{-6}$ in Eq. (3.12).
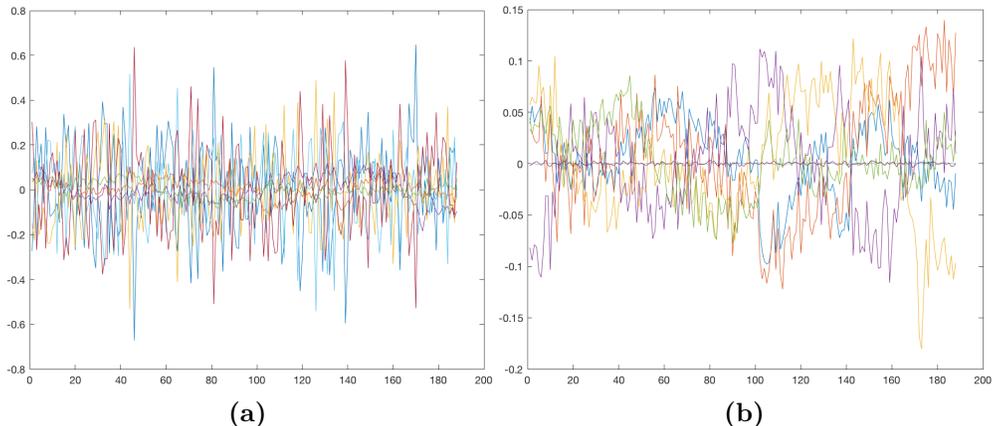
Figure 3.10: Estimated $\mathbf{W}_1$ when $c = 7$ (ground truth $c = 5$ for the generated data) given noisy data with SNR $= 20$ dB. (a) Estimated $\mathbf{W}_1$ without l_{21} constraint. (b) Estimated $\mathbf{W}_1$ with l_{21} constraint.

Figure 3.10 shows the results of the estimated $\mathbf{W}_1$ when the number of endmembers is set to $c = 7$ while the ground truth is $c = 5$. We can observe that, with the l_{21} constraint on $\mathbf{W}_1$, the redundant endmembers are almost zeros as shown in Fig. 3.10b, because the l_{21} constraint encourages the rows of $\mathbf{W}_1$ to be zeros, while at the mean time, the reconstruction term $\mathbf{X} - \mathbf{AS}$ is encouraged to be small.
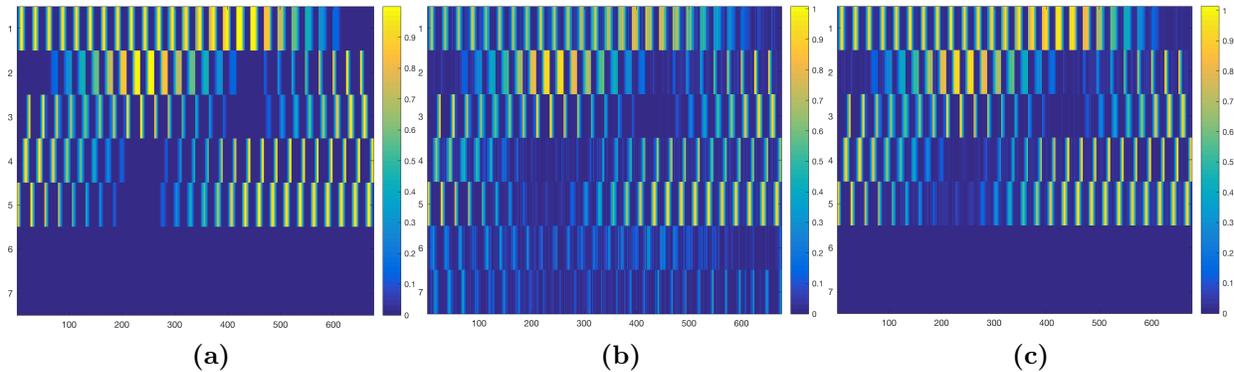
Figure 3.11: Estimated $\mathbf{S}$ given noisy data with SNR $=20$ dB and $c = 5$. (a) Ground truth abundance $\mathbf{S}$. (b) Estimated $\mathbf{S}$ without the l_{21} constraint given $c = 7$. AID $= 6.08$. (c) Estimated $\mathbf{S}$ with the l_{21} constraint given $c = 7$. AID $= 0.81$.

The zero rows in $\mathbf{W}_1$ result in the de-activation of the hidden nodes, abundance $\mathbf{S}$, in the network. In Fig. 3.11, the top image shows the ground truth $\mathbf{S} \in \mathbb{R}^{5 \times 1024}$ with pseudocolor. The columns of the image represent the 1024 pixels and the rows of the image denotes the c hidden nodes. For comparison, we augment the rows of $\mathbf{S}$ to be 7, with 2 additional zero rows. Without the l_{21} constraint, the method yields the endmember as shown in Fig. 3.11b, which is very different from the ground truth abundance $\mathbf{S}$ with AID $= 6.08$. But the estimated $\mathbf{S}$ after applying the l_{21} constraint is very close to the ground truth, as shown in Fig. 3.11c. And the AID is largely reduced to 0.81.

To further show the effectiveness of this constraint, we increase the number of endmembers c from 2 to 10 for the input data with noise level from SNR $= 50$ dB to 10 dB. The parameter γ in Eq. (3.12) is increased to $\gamma = 10^{-3}$ in this group of experiments to better demonstrate the influences of the l_{21} norm. The reconstruction errors of the network given different c are drawn in Fig. 3.12. The results show that when the number of endmembers is increased but smaller than the ground truth $c = 5$, the reconstruction error is very low. However, when the reconstruction error reaches the lowest value at $c = 5$, it begins to increase noticeably when c is increased. In all the experiments, we are able to find the correct number of endmembers c no matter what the level of noise is in the data.
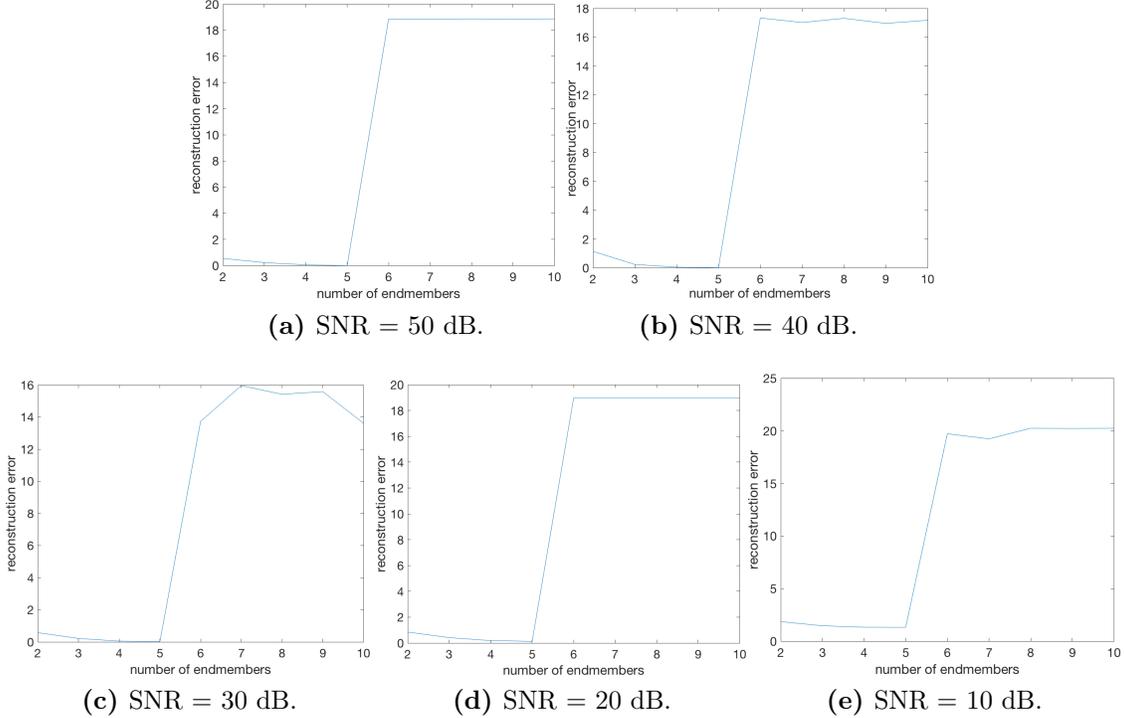
Figure 3.12: Estimated number of endmembers c given noisy data with SNR = 50 dB to 10 dB.

Comparison to State-of-the-Art Unmixing Algorithms:

In this experiment, we illustrate the performance of the proposed method with both the denoising constraint $\mathbf{W}_n$ on the decoder $\mathbf{A}$ and the l_{21} constraint on the encoder $\mathbf{W}_1$, by comparing it with six state-of-the-art algorithms. Table 3.3 reports the performance of different unmixing algorithms with different noise levels from SNR = 50 dB to 10 dB. In all the experiments, we assume that $c = 5$ is known. Note that, we perform each algorithm ten times on each dataset and reports the average value in Table 3.3.

As demonstrated in Table 3.3, when the noise level is relatively low (high SNR), most approaches achieve very good results, but the performance of unmixing drops when the noise level is increased (low SNR). The Bayesian based approach [38] achieves the best AAD when SNR = 40 dB. But the performance decreases when SNR > 30 dB. The geometric-based approach MVC-NMF [109] achieves better results than MVSA [94] although their main principle are the same, i.e., to find the endmembers that circumscribing the data and embedding the minimum volume. This is because MVC-NMF applies the minimum volume
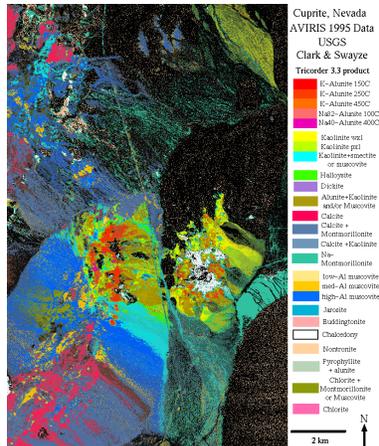
Figure 3.13: Different minerals in the Cuprite data.

constraint directly on the endmembers, while MVSA applied it on the abundance. The most stable approaches are MLNMF [122], R-CoNMF [95], and the proposed uDAS. The benefit of MLNMF [122] is that it solves the problem through a multilayer-NMF, which encourages the abundance in each layer to be sparse, which is more reasonable than forcing the whole abundance to be sparse. R-CoNMF [95] achieves better results by projecting the data onto a lower space and forcing its endmembers to be close to the extracted endmembers from VCA, while maintaining a low reconstruction error. The proposed algorithm outperforms the other approaches and achieves the best or comparable results in all cases, especially when the noise level is high. This experiment well indicate that the proposed uDAS is robust to noise.

3.6.4 Experiments with Real Data

There is no agreement on the number of endmembers in Cuprite by recent literature [12, 94, 95, 109, 122, 72, 64, 134]. But some endmembers are duplicated when the number of endmembers is set to large. We set the number of endmembers as $c = 12$ in this experiment. The extracted endmembers along with the corresponding USGS library signatures are shown in Fig. 3.14. Both the visualized results and the calculated SAD show the effectiveness of the proposed approach. For demonstration purpose, the estimated abundance maps are presented in Fig. 3.15.
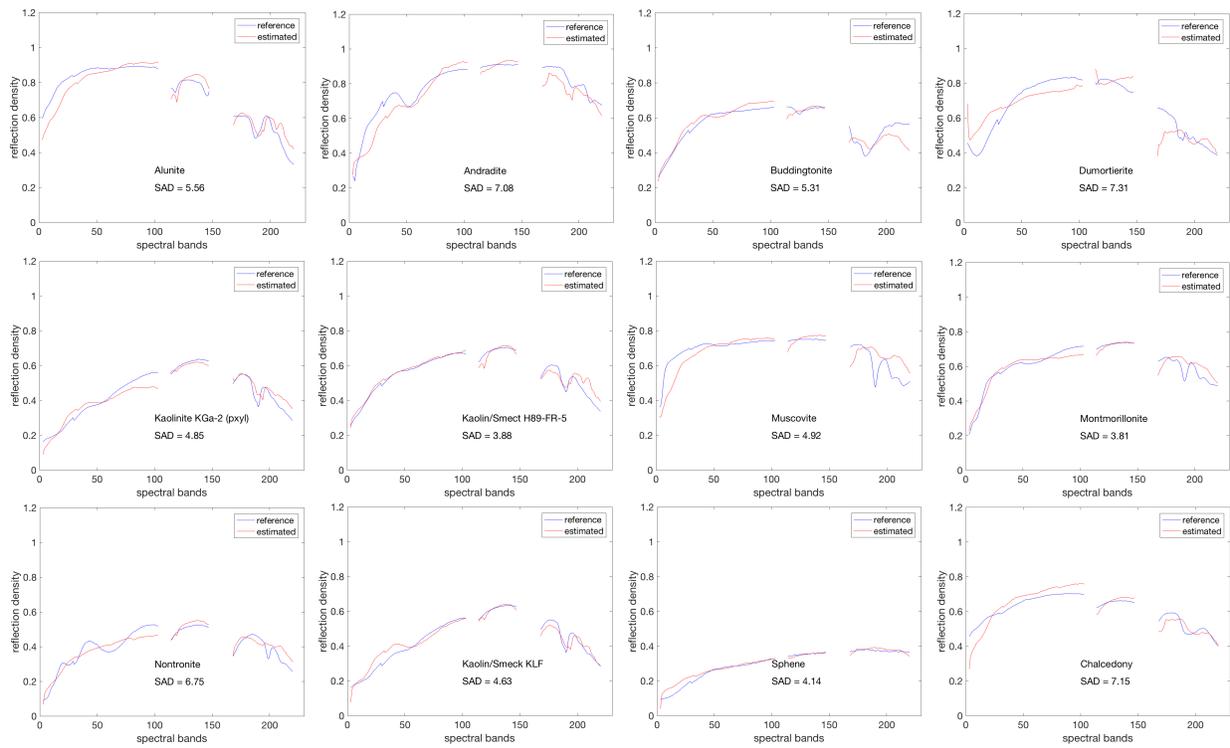
Figure 3.14: Extracted endmembers of the AVIRIS Cuprite data from the proposed uDAS.

Table 3.1: Notations

Symbol	Description
$\mathbf{X}$	input noise-corrupted data
$\hat{\mathbf{X}}$	reconstructed data, reconstructed clean data
$\mathbf{W}_1$	encoder weights of the network
$\mathbf{S}$	abundance, hidden layer output
$\mathbf{A}$	endmembers, decoder weights of the network
σ	activation function
$\mathbf{N}$	noise
c	the number of endmembers
$\mathbf{I}$	identity matrix
$\mathbf{W}_n$	denoising constraints on the decoder
$\tilde{\mathbf{X}}$	stacked different corrupted data $\mathbf{X}$
$\mathbf{X_r}$	the repeated clean data $\hat{\mathbf{X}}$
p	the corruption probability
$\bar{\mathbf{X}}$	augmented $\mathbf{X}$ defined in Eq. (3.10)
$\bar{\mathbf{A}}$	augmented $\mathbf{A}$ defined in Eq. (3.10)
$\mathbf{b}$	bias
l_{21}	the sum of the Euclidean norms of the columns of the matrix.
λ	trade off between the reconstruction error and denoising constraint
γ	trade off between the reconstruction error and l_{21} sparsity constraint

Table 3.2: Denoising constraint evaluation on synthetic data with SNR = 20 dB and 10 dB.

SNR = 20 dB	SAD	SID
without denoising constraint	1.63	0.003
with cascaded denoising network	1.53	0.0023
with the denoising constraint $\mathbf{W}_n$	**1.12**	**0.001**
SNR = 10 dB	SAD	SID
without denoising constraint	6.09	0.09
with cascaded denoising network	5.71	0.0939
with the denoising constraint $\mathbf{W}_n$	**4.56**	**0.025**

Table 3.3: Unmixing evaluation on synthetic data with Gaussian noise from SNR = 50 dB to 10 dB.

SNR = 50 dB	RMSE	AAD	AID	SAD	SID
MVC-NMF	1.6e-2	2.98	0.350	3.24	5.9e-2
Bayesian	1.97e-5	0.375	2.71e-2	0.349	1.34e-4
MVSA	0.322	79.2	20.3	33.6	0.504
MLNMF	3.94e-5	0.663	**4.60e-3**	1.59	1.89e-3
R-CoNMF	4.16e-6	0.181	3.65e-2	0.132	1.97e-5
Cascaded-uDAS	5.63e-5	0.631	0.131	0.105	1.16e-5
uDAS	**1.18e-6**	**0.113**	1.42e-2	**0.104**	**6.76e-6**
SNR = 40 dB	RMSE	AAD	AID	SAD	SID
MVC-NMF	2.48e-3	3.25	0.329	2.95	5.28e-3
Bayesian	1.55e-5	**0.330**	6.15e-2	0.141	1.90e-5
MVSA	0.271	64.8	14.0	2.95	5.28e-3
MLNMF	3.36e-5	0.625	6.87e-2	1.77	5.77e-3
R-CoNMF	4.64e-5	0.578	0.142	0.433	1.97e-4
Cascaded-uDAS	1.97e-4	0.951	0.215	0.181	2.67e-5
uDAS	**1.15e-5**	0.332	**5.51e-2**	**0.106**	**7.50e-6**
SNR = 30 dB	RMSE	AAD	AID	SAD	SID
MVC-NMF	2.52e-3	4.11	0.491	3.42	6.10e-3
Bayesian	8.86e-3	6.39	1.33	3.13	8.59e-3
MVSA	0.287	76.9	15.9	3.42	6.10e-3
MLNMF	1.22e-4	1.2	0.236	1.76	3.98e-3
R-CoNMF	3.50e-4	1.77	0.401	1.30	1.99e-3
Cascaded-uDAS	7.61e-4	2.41	0.477	0.481	2.27e-4
uDAS	**1.18e-4**	**1.17**	**0.189**	**0.363**	**1.61e-4**
SNR = 20 dB	RMSE	AAD	AID	SAD	SID
MVC-NMF	2.1e-2	3.46	0.738	2.90	5.2e-3
Bayesian	5.40e-3	4.98	1.43	4.03	2.79e-2
MVSA	0.175	49.6	9.32	2.89	5.16e-3
MLNMF	1.24e-3	4.12	0.742	2.19	8.28e-3
R-CoNMF	1.89e-3	3.99	0.957	1.55	2.26e-3
Cascaded-uDAS	3.9e-3	5.52	1.34	1.53	2.3e-3
uDAS	**1.20e-3**	**3.32**	**0.69**	**1.12**	**1.0e-3**
SNR = 10 dB	RMSE	AAD	AID	SAD	SID
MVC-NMF	0.188	49.4	12.6	11.2	7.96e-2
Bayesian	1.86e-2	11.2	2.68	6.22	3.95e-2
MVSA	0.212	66.1	12.7	11.2	7.96e-2
MLNMF	2.48e-2	13.5	3.05	6.72	4.89e-2
R-CoNMF	1.21e-2	11.5	2.27	5.04	3.16e-2
Cascaded-uDAS	1.5e-2	12.3	2.57	5.71	9.39e-2
uDAS	**9e-3**	**10.3**	**2.05**	**4.56**	**2.59e-2**
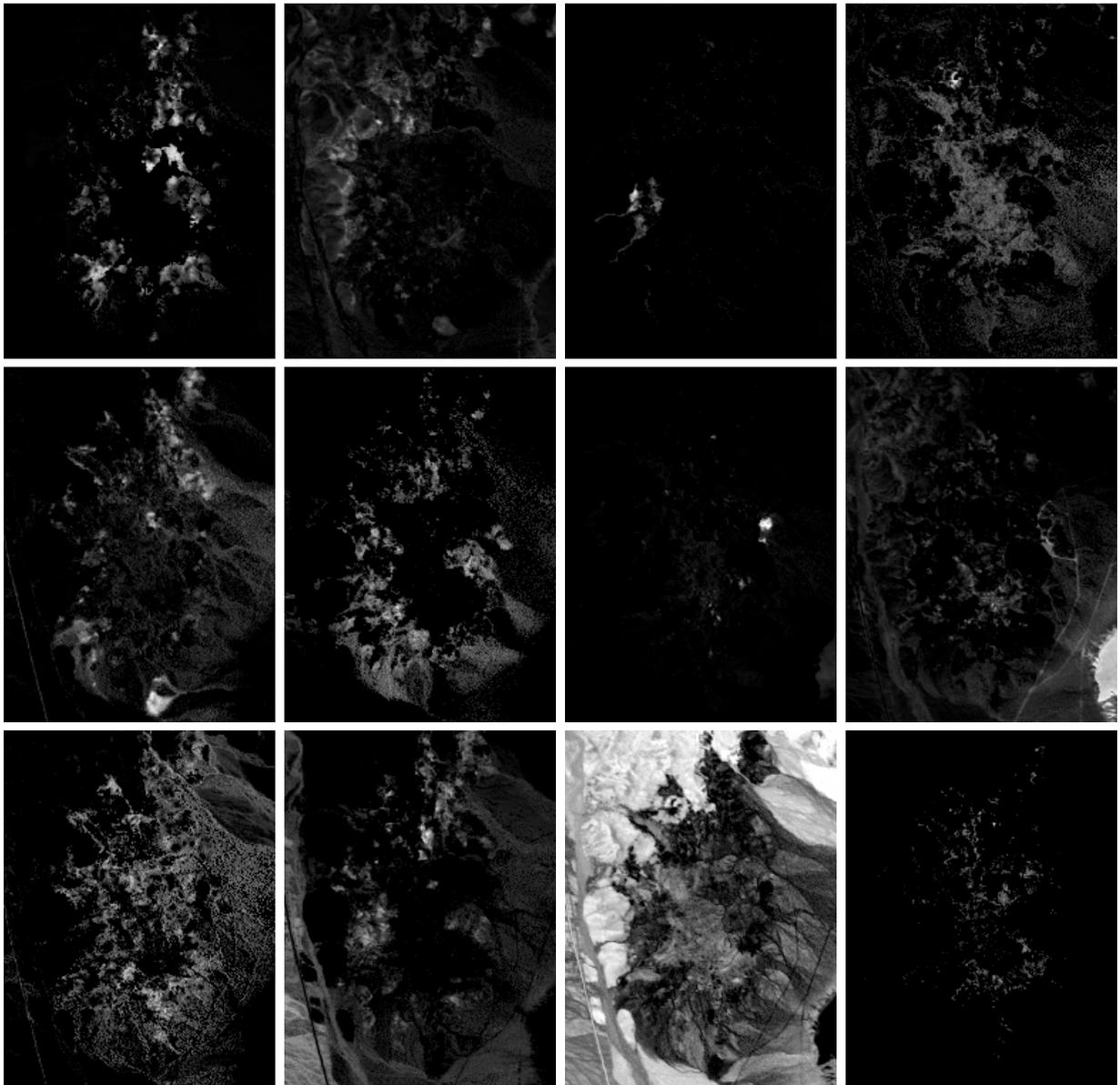
Figure 3.15: Estimated abundance maps of the AVIRIS Cuprite data corresponding to the Fig 3.14.

3.6.5 Parameter Setting

There are three free parameters in the proposed method, the noise level p in mDA, the parameter λ that balances the decoding weights $\mathbf{A}$ and the parameter γ that balances the encoding weights $\mathbf{W}_1$. Through extensive unmixing experiments, we set p to be as small as 0.00001, γ to be 10^{-6}, and λ to be 10^{-8}.

Chapter 4

Hyperspectral Image Super-Resolution

4.1 Introduction

Compared to multispectral images (MSI) or conventional color images (RGB) with much less number of spectral bands, HSI collects hundreds of contiguous bands which provides rich details to distinguish different objects. However, its spatial resolution is significantly lower than MSI or RGB due to hardware limitations [77, 4]. On the contrary, although MSI or RGB has high spatial resolution, their spectral resolution is relatively low. Very often, to yield better recognition and analysis results, images with both high spectral and spatial resolution are desired [152]. A natural way to generate such images is to fuse hyperspectral images with multispectral images or conventional color images. This procedure is referred to as *hyperspectral image super-resolution (HSI-SR)* [4, 87, 37] as shown in Fig. 4.1.

The problem of HSI-SR originates from *multispectral pan-sharpening (MSI-PAN)* in the remote sensing field, where the spatial resolution of MSI is further improved by a high-resolution panchromatic image (PAN). Note that, in general, resolution refers to the spatial resolution. Usually, MSI has much higher resolution than HSI, but PAN has even higher resolution than MSI. We use LR to denote low spatial resolution and HR for high spatial resolution. There are roughly two groups of MSI-PAN methods, namely, the component substitution (CS) [145, 25, 3] and the multi-resolution analysis (MRA) based approaches [2].
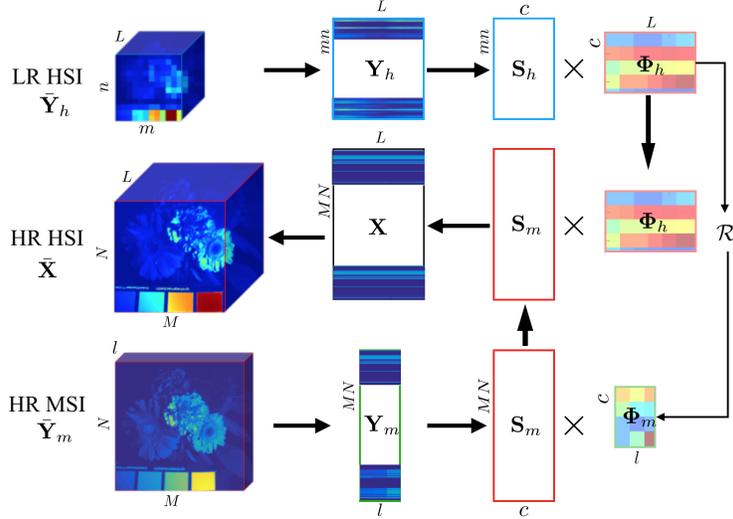
Figure 4.1: General procedure of HSI-SR.

Although MSI-PAN has been well developed through decades of innovations [145, 100, 179], they cannot be readily adopted to solve the HSI-SR problems. On one hand, the amount of spectral information to be preserved for HSI-SR is much higher than that of MSI-PAN, thus it is easier to introduce spectral distortion, *i.e.*, the output image does not preserve the accurate spectral information [100, 165, 4, 37]. On the other hand, HSI possesses much lower resolution than that of MSI, making it more challenging to improve the spatial resolution. There have been few methods specifically designed for HSI-SR. But existing HSI-SR approaches [100, 179, 87, 165] generally assume that the LR HSI is down-sampled from the HR HSI and such down-sampling function is used as a prior in the fusion methods, which may not be true in practice due to the distortions caused by both the sensors and complex environmental conditions [4].

HSI-SR is also closely related to the natural image super-resolution (SR) problem, which has been extensively studied and achieved excellent performance through the state-of-the-art *deep learning* [42, 102, 132, 81, 82, 91, 86, 62]. The main principle of SR is to learn a mapping function between LR images and HR images in a supervised fashion. Natural image SR methods usually work on up to 4× upscaling. There have been three attempts to address the MSI-PAN problem with deep learning where the mapping function is learned using different frameworks including tied-weights denoising/ autoencoder [70], SRCNN [107],

47
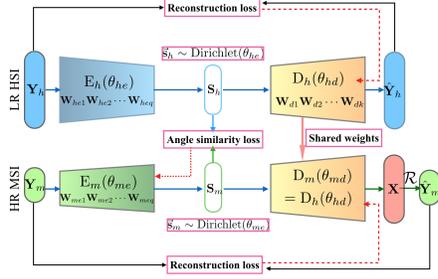
Figure 4.2: Simplified architecture of the proposed uSDN.
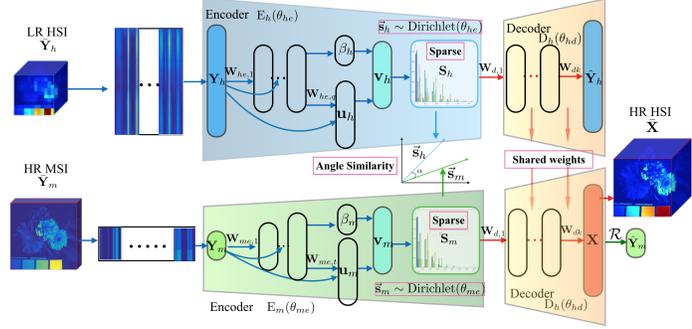
Figure 4.3: Details of the encoder nets.

and deep residual network [62, 158]. These deep learning based methods, including natural image SR and MSI-PAN are all supervised, making their adoption on HSI-SR a challenge due to two reasons. First, they are designed to find an end-to-end mapping function between the LR images and HR images under the assumption that the mapping function is the same for different images. However, the mapping function may not be the same for images acquired with different sensors. Even for the data collected from the same sensor, the mapping function for different spectral bands may not be the same. Thus the assumption may cause severe spectral distortion. Second, training a mapping function is a supervised solution which requires a large dataset, the down-sampling function, and the availability of the HR HSI, that are not realistic for HSI.

To overcome the drawbacks of existing approaches, we propose an *unsupervised* network structure to address the challenges of HSI-SR. To the best of our knowledge, this is the first effort to solving the HSI-SR problem with deep learning in an unsupervised fashion. The proposed method is referred to as uSDN.

4.2 Problem Formulation

Given the LR HSI, $\bar{\mathbf{Y}}_h \in \mathbb{R}^{m \times n \times L}$, where m, n and L denote the width, height and number of spectral bands of the HSI, respectively, and the corresponding HR MSI, $\bar{\mathbf{Y}}_m \in \mathbb{R}^{M \times N \times l}$, where M, N and l denote the width, height and number of spectral bands of the MSI, respectively, the goal is to estimate the HR HSI, $\bar{\mathbf{X}} \in \mathbb{R}^{M \times N \times L}$, with both high spatial

and spectral resolution. In general, MSI has much higher spatial resolution than HSI, *i.e.*, $M \gg m$, $N \gg n$, and HSI has much higher spectral resolution than MSI, *i.e.*, $L \gg l$. To facilitate the subsequent processing, we unfold the 3D images into 2D matrices, *i.e.*, each row of the 2D matrix denotes the spectral reflectance of a given pixel. The unfolded matrices are written as $\mathbf{Y}_h \in \mathbb{R}^{mn \times L}$, $\mathbf{Y}_m \in \mathbb{R}^{MN \times l}$ and $\mathbf{X} \in \mathbb{R}^{MN \times L}$. This is illustrated in Fig. 4.1.

Assuming that each row of $\mathbf{Y}_h$ is a linear combination of c basis vectors (or spectral signatures), as expressed in Eq. (4.1), where $\mathbf{\Phi}_h \in \mathbb{R}^{c \times L}$ and each row of which denotes the spectral basis that preserves the spectral information and $\mathbf{S}_h \in \mathbb{R}^{mn \times c}$ is the corresponding proportional coefficients (referred to as *representations* in deep learning). Since the coefficients indicate how the spectral bases are mixed at specific spatial locations, they preserve the spatial structure of HSI.

Similarly, $\mathbf{Y}_m$ can be expressed as Eq. (4.2), where $\mathbf{\Phi}_m \in \mathbb{R}^{c \times l}$ and each row of which indicates the spectral basis of MSI. $\mathcal{R} \in \mathbb{R}^{L \times l}$ is the transformation matrix given as a prior from the sensor [77, 166, 156, 100, 135, 152, 87, 37], which describes the relationship between HSI and MSI bases. With $\mathbf{\Phi}_h \in \mathbb{R}^{c \times L}$ carrying the high spectral information and $\mathbf{S}_m \in \mathbb{R}^{MN \times c}$ carrying the high spatial information, the desired HR HSI, $\mathbf{X}$, is generated by Eq. (4.3). See Fig. 4.1.

$$\mathbf{Y}_h = \mathbf{S}_h \mathbf{\Phi}_h, \tag{4.1}$$

$$\mathbf{Y}_m = \mathbf{S}_m \mathbf{\Phi}_m, \quad \mathbf{\Phi}_m = \mathbf{\Phi}_h \mathcal{R} \tag{4.2}$$

$$\mathbf{X} = \mathbf{S}_m \mathbf{\Phi}_h. \tag{4.3}$$

The problem of HSI-SR can be described mathematically as $P(\mathbf{X}|\mathbf{Y}_h, \mathbf{Y}_m)$. Since the ground truth $\mathbf{X}$ is not available, the problem should be solved in an unsupervised fashion. The key to addressing this problem is to take advantage of the shared information, *i.e.*, $\mathbf{\Phi}_h \in \mathbb{R}^{c \times L}$, to extract desired high spectral bases $\mathbf{\Phi}_h$ and spatial representations $\mathbf{S}_m$ from two different modalities.

In addition, three unique requirements of HSI-SR need to be given special consideration. First, in representing HSI or MSI as a linear combination of spectral signatures, the

representation vectors should be non-negative and sum-to-one. That is, $\sum_{j=1}^{c} s_{ij} = 1$, where $\mathbf{s}_i$ is the row vector of either $\mathbf{S}_h$ or $\mathbf{S}_m$ [77, 166, 43, 87, 150]. Second, due to the fact that each pixel of image only consists of a few spectral bases, the representations should be sparse. Third, spectral distortion should be largely reduced in the process in order to preserve the spectral information of HR HSI while gaining spatial resolution.

4.3   Proposed Approach

We propose an unsupervised architecture as shown in Fig. 4.2. We highlight the three structural uniquenesses here. First, the architecture consists of two deep networks, for the representation learning of the LR HSI and HR MSI, respectively. These two networks share the same decoder weights, enabling the extraction of both spectral and spatial information from multi-modalities in an unsupervised fashion. Second, in order to satisfy the sum-to-one constraint of the representations, both $\mathbf{S}_h$ and $\mathbf{S}_m$ are encouraged to follow a Dirichlet distribution where the sum-to-one property is naturally incorporated in the network with a further sparsity constraint. Third, to address the challenge of spectral distortion, the representations of two modalities are encouraged to have similar patterns by minimizing their angular difference.

4.3.1   Network Architecture

As shown in Fig. 4.2, the network reconstructs both the LR HSI $\mathbf{Y}_h$ and HR MSI $\mathbf{Y}_m$ in a coupled fashion. Taking the LR HSI network (the top network) as an example. The network consists of an encoder $\mathrm{E}_h(\theta_{he})$, which maps the input data to low-dimensional representations (latent variables on the Bottleneck hidden layer), $i.e.$, $p_{\theta_{he}}(\mathbf{S}_h|\mathbf{Y}_h)$, and a decoder $\mathrm{D}_h(\theta_{hd})$ which reconstructs the data from the representations, $i.e.$, $p_{\theta_{hd}}(\hat{\mathbf{Y}}_h|\mathbf{S}_h)$. Both the encoder and decoder are constructed with multiple fully-connected layers. Note that the bottleneck hidden layer $\mathbf{S}_h$ behaves as the representation layer that reflect the spatial information and the weights θ_{hd} of the decoder $\mathrm{D}_h(\theta_{hd})$ serve as $\mathbf{\Phi}_h$ in Eq. (4.1), respectively. This correspondence is further elaborated below.

The HSI is reconstructed by $\hat{\mathbf{Y}}_h = f_k(\mathbf{W}_{dk}f_{k-1}(...(f_1(\mathbf{S}_h\mathbf{W}_{d1}+b_1)...)+b_{k-1})+b_k)$, where $\mathbf{W}_{dk}$ denotes the weights in the kth layer. To extract the spectral basis from LR HSI, the latent variables of the representation layer $\mathbf{S}_h$ act as the proportional coefficients, where $\mathbf{S}_h$ follows a Dirichlet distribution with the sum-to-one property naturally incorporated. Suppose the activation function is an identity function and there is no bias in the decoder, we have $\theta_{hd} = \mathbf{W}_1\mathbf{W}_2...\mathbf{W}_k$. That is, the weights θ_{hd} of the decoder correspond to the spectral basis $\mathbf{\Phi}_h$ in Eq. (4.1) and $\mathbf{\Phi}_h = \theta_{hd}$. In this way, $\mathbf{\Phi}_h$ preserves the spectral information of LR HSI, and the latent variables $\mathbf{S}_h$ preserves the spatial information effectively.

Equivalently, the bottom network reconstructs the HR MSI in a similar way with encoder $\mathrm{E}_m(\theta_{me})$ and decoder $\mathrm{D}_m(\theta_{md})$. However, since $l \leq c \leq L$, $i.e.$, the number of latent variables, L, is much larger than the number of input nodes, l, the MSI network is very unstable and hard to train. On the other hand, the spectral basis of HR MSI can be transformed from those of LR HSI which possesses more spectral information, the decoder of the MSI is designed to share the weights with that of HSI in terms of $\theta_{md} = \mathbf{\Phi}_m = \theta_{hd}\mathcal{R} = \mathbf{\Phi}_h\mathcal{R}$. Then the reconstructed HR MSI can be obtained by $\hat{\mathbf{Y}}_m = \mathbf{S}_m\mathbf{\Phi}_h\mathcal{R}$. In this way, only the encoder $\mathrm{E}_m(\theta_{me})$ of the MSI is updated during the optimization, where the HR spatial information $\mathbf{S}_m$ is extracted from MSI. Eventually, the desired HR HSI is generated directly by $\mathbf{X} = \mathbf{S}_m\mathbf{\Phi}_h$. Note that the dashed lines in the image show the path of backpropagation which will be elaborated in Sec. 4.3.4.

4.3.2   Sparse Dirichlet-Net with Dense Connectivity

To extract stable spectral information, we need to enforce the proportional coefficients $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_i, \cdots, \mathbf{s}_p)^T$ of each pixel to sum-to-one [166, 161, 87, 87], $i.e.$, $\sum_{j=1}^c s_{ij} = 1$. Without loss of generality, $\mathbf{S}$ represents either $\mathbf{S}_h$ with $p = mn$ or $\mathbf{S}_m$ with $p = MN$. In addition, due to the fact that only a few spectral bases actually contribute in the linear combination of the spectral reflectance of each pixel, the coefficients should also be sparse. In the proposed architecture, the latent variables (or representations) of the hidden layer $\mathbf{S}_h$ or $\mathbf{S}_m$ correspond to the proportional coefficients in Eqs. (4.1) and (4.2). To naturally incorporate the sum-to-one property, the representations are encouraged to follow a Dirichlet

distribution which is accomplished with stick-breaking process as illustrated in Fig. 3. Furthermore, entropy function is adopted to reinforce the sparsity of the representations.

The stick-breaking process was first proposed by Sethuranman [130] back in 1994. It is used to generate random vectors $\mathbf{s}$ with Dirichlet distribution. The process can be illustrated as breaking a unit-length stick into c pieces, the length of which follows a Dirichlet distribution. Assuming that the generated vector is denoted as $\mathbf{s} = (s_1, \cdots, s_j, \cdots, s_c)$, we have $0 \leq s_j \leq 1$, and the variables in the vector are sum to one, *i.e.*, $\sum_{j=1}^{c} s_j = 1$. Mathematically [130], a single variable s_j is defined as

$$s_j = \begin{cases} v_1 & \text{for} \quad j = 1 \\ v_j \prod_{o<j}(1 - v_o) & \text{for} \quad j > 1, \end{cases} \tag{4.4}$$

where v_o is drawn from a Beta distribution, *i.e.*, $v_o \sim \text{Beta}(u, \alpha, \beta)$. Nalisnick and Smyth successfully coupled the expressiveness of generative networks with Bayesian nonparametric model through stick-breaking process [111]. The network uses a Kumaraswamy distribution [83] as an approximate posterior which takes in the samples from a randomly generated uniform distribution during the training procedure.

Different from the generative network, we aim to find shared representations that better reconstruct the data. Therefore, the weights of the network should be changed according to the input data instead of randomly generated distribution. It has been proved that when $v_o \sim \text{Beta}(u, 1, \beta)$, $\mathbf{s}$ follows a Dirichlet distribution. Since it is difficult to draw samples directly from Beta distribution, we draw samples from the inverse transform of Kumaraswamy distribution, as shown in Eq. (4.5), which is equivalent to Beta distribution when $\alpha = 1$ or $\beta = 1$,

$$\text{kuma}(u, \alpha, \beta) = \alpha \beta u^{\alpha-1}(1 - u^\alpha)^{\beta-1} \tag{4.5}$$

where $\alpha > 0$, $\beta > 0$ and $u \in (0, 1)$. The benefit of Kumaraswamy distribution is that it has a closed-form CDF, where the inverse transform is defined as

$$v_o \sim (1 - (1 - u^{\frac{1}{\beta}})^{\frac{1}{\alpha}}). \tag{4.6}$$
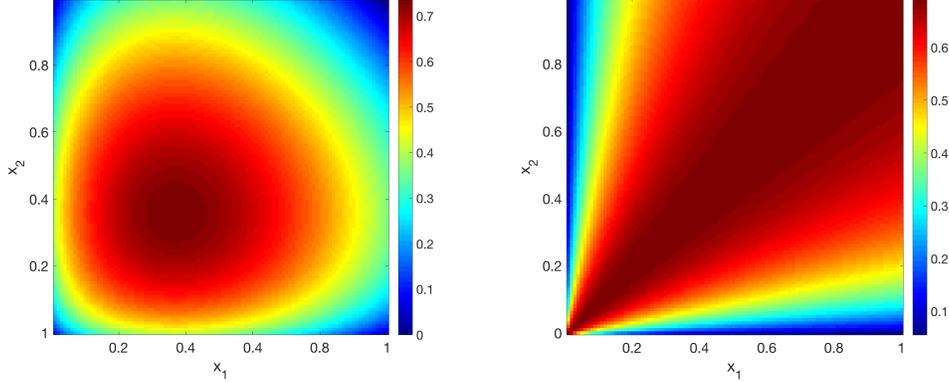
Figure 4.4: Shannon entropy (L) and Shannon entropy function (R).

Let $\alpha = 1$, parameters u and β are learned through the network as illustrated in Fig. 4.3. Because $\beta > 0$, a softplus is adopted as the activation function [46] at the β layer. Similarly, a sigmoid [59] is used to map u into $(0, 1)$ range at the **u** layer. To avoid gradient vanishing and increase the representation power of the proposed method, the encoder of the network is densely connected, *i.e.*, each layer is fully connected with all its subsequent layers [68].

To further increase the variability of u and β (theoretically, we want the learned u and β to be any number within their range), instead of concatenating all the preceding layers, the input of the kth layer is the summation of all the preceding layers x_0, x_1, x_{k-1} with their own weights, *i.e.*, $\mathbf{W}_0 x_0 + \mathbf{W}_1 x_1 + ... + \mathbf{W}_{k-1} x_{k-1}$. In this way, fewer number of layers is required to learn the optimal representations.

Although the stick-breaking structure encourages the representations to follow a Dirichlet distribution, it does not guarantee the sparsity of the representations. In addition, the widely used l_1 regularization or Kullback-Leibler divergence [55] will not encourage the representation layer to be sparse either, because they guarantee the sparsity by reducing the mean of active value, *i.e.*, mean of the representation layer. However, due to the stick-breaking structure, the mean of $\mathbf{S}_h$ or $\mathbf{S}_m$ is almost one. Therefore, we introduce a generalized Shannon entropy function [69] to reinforce the sparsity of the representation layer which works effectively even with the sum-to-one constraint.

The entropy function was first proposed in compressive sensing field to solve the signal recovery problem. It is defined as

$$\mathcal{H}_p(\mathbf{s}) = -\sum_{j=1}^{N} \frac{|s_j|^p}{\|\mathbf{s}\|_p^p} \log \frac{|s_j|^p}{\|\mathbf{s}\|_p^p}. \tag{4.7}$$

Compared to the more popular Shannon entropy, the entropy function Eq. (4.7) decreases monotonically when the data become sparse. To illustrate the effect, we show the phenomena with 2D variables in Fig. 4.4. Shannon entropy is small when both x_1 and x_2 are small or large. But for Shannon entropy function, the local minimum only occurs at the boundaries of the quadrants. This nice property guarantees the sparsity of arbitrary data even the data are with the sum-to-one constraint. Due to the stick-breaking structure, the latent variables at the representation layer are positive. We choose $p = 1$ which is more efficient and will encourage the variables to be sparse.

4.3.3 Angle Similarity

Extracting spatial information from HR MSI is quite challenging and easy to introduce spectral distortion in the subsequent HR HSI results. The main cause to this problem is that the number of the representations c (number of nodes in the representation layer) is much larger than the dimension of the MSI, $i.e.$, $c \gg l$. Previous researchers assume the down-sampling function is available a-$priori$ to build a relationship between the representations of HSI and MSI. However, the down-sampling function is usually unknown for real applications.

Therefore, instead of taking the down-sampling function as a prior, we encourage the representations $\mathbf{S}_h$ and $\mathbf{S}_m$ of the two networks following a similar pattern to prevent spectral distortion. And such similarity is measured by the angular difference between the two representations. Spectral angle mapper (SAM) is employed to measure this angular difference. SAM is a spectral evaluation method in remote sensing [100, 165, 118], which measures the angular difference between the estimated image and the ground truth image. The lower the SAM score, the smaller the spectral angle difference,, and the more similar the two representations.

Since the HSI and MSI networks share the same decoder weights, the representations should have similar angle in order to generate high quality image with less spectral distortion. Besides encouraging the representation layer to follow a sparse Dirichlet distribution, we further reduce the angular difference of the representations of HSI and MSI during the optimization procedure.

In the network, representations $\mathbf{S}_h \in \mathbb{R}^{mn \times c}$ and $\mathbf{S}_m \in \mathbb{R}^{MN \times c}$, from two different modalities have different dimensions. To minimize the angular difference, we increase the size of the low-dimensional $\mathbf{S}_h$ by duplicating its values at each pixel to its nearest neighborhood. Then the duplicated representations $\tilde{\mathbf{S}}_h \in \mathbb{R}^{MN \times c}$ have the same dimension as $\mathbf{S}_m$. With vectors of equal size, the angular difference is defined as

$$\mathcal{A}(\tilde{\mathbf{S}}_h, \mathbf{S}_m) = \frac{1}{MN} \sum_{i=1}^{MN} \arccos\left(\frac{\tilde{\mathbf{s}}_h^i \cdot \mathbf{s}_m^i}{\|\tilde{\mathbf{s}}_h^i\|_2 \|\mathbf{s}_m^i\|_2}\right) \tag{4.8}$$

To map the range of the angle within $(0, 1)$, Eq. (4.8) is divided by the circular constant π.

$$\mathcal{J}(\tilde{\mathbf{S}}_h, \mathbf{S}_m) = \frac{\mathcal{A}(\tilde{\mathbf{S}}_h, \mathbf{S}_m)}{\pi} \tag{4.9}$$

4.3.4 Optimization and Implementation Details

To prevent over-fitting, we applied an l_2 norm on the decoder weights. The objective functions of the proposed network architecture can then be expressed as:

$$\mathcal{L}(\theta_{he}, \theta_{hd}) = \frac{1}{2} \|\mathbf{Y}_h(\theta_{he}, \theta_{hd}) - \hat{\mathbf{Y}}_h(\theta_{he}, \theta_{hd})\|_F^2 \\ + \lambda \mathcal{H}_1(\mathbf{S}_h(\theta_{he})) + \mu \|\theta_{hd}\|_F^2, \tag{4.10}$$

$$\mathcal{L}(\theta_{me}) = \frac{1}{2} \|\mathbf{Y}_m(\theta_{me}, \theta_{hd}) - \hat{\mathbf{Y}}_m(\theta_{me}, \theta_{hd})\|_F^2 \\ + \lambda \mathcal{H}_1(\mathbf{S}_m(\theta_{me})), \tag{4.11}$$

$$\mathcal{L}(\theta_{me}) = \mathcal{J}(\tilde{\mathbf{S}}_h(\theta_{he}), \mathbf{S}_m(\theta_{me})), \tag{4.12}$$

where λ and μ are parameters that balance the trade-off between the reconstruction error and the sparsity and weights loss, respectively.

The proposed architecture consists of two sparse Dirichlet-Nets which extract the spectral information Φ_h from HSI and spatial information $\mathbf{S}_m$ from MSI. The network is optimized with back-propagation following the procedure described below, also illustrated in Fig. 4.2 with the dashed line.

Step 1: Since the decoder weights θ_{hd} of the HSI network preserves the spectral information Φ_h, we first update the HSI network, given the objective function in Eq. (4.10), to find the optimal θ_{hd}. To prevent over-fitting, an l_2 norm is applied on the decoder of the HSI network.

Step 2: The estimated decoder weights θ_{hd} are fixed and shared with the decoder of the MSI network. Update the encoder weights θ_{me} of the MSI network given the objective function in Eq. (4.11).

Step 3: To reduce spectral distortion, every 10 iterations, we minimize the angular difference between the representations of two modalities given the objective function in Eq. (4.12). Since we already have θ_{he} from the first step, only the encoder θ_{me} of the MSI network is updated during the optimization.

Table 4.1: The number of layers and nodes in the network.

Dirichlet-Net	Encoder			
	#layers and #nodes	$\mathbf{u}$	β	$\mathbf{v}$
HSI	3 / [10,10,10]	10	1	10
MSI	5 / [4,5,7,910]	10	1	10

For all the experiments, both the input and output of the HSI network have 31 nodes, representing the number of spectral bands in the data. The numbers of densely-connected layers and nodes of the encoder are shown in Table 4.1. There are 3 layers in the HSI network and each layer contains 10 nodes. The MSI network has 5 layers with the number of nodes increases from 4 to 10. The $\mathbf{v}_h/\mathbf{v}_m$ are drawn with Eq. (4.6) given $\mathbf{u}_h/\mathbf{u}_m$ and β_h/β_m, which are learned by back-propagation. Both β_h and β_m have only one node, denoting the distribution parameter of each pixel. The representation layers, $\mathbf{S}_h$ and $\mathbf{S}_m$ with 10 nodes are constructed with $\mathbf{v}_h$ and $\mathbf{v}_m$, respectively, according to Eq. (4.4). The network shares the decoder with 2 layers and each layer has 10 nodes.

4.4 Experiments and Results

4.4.1 Datesets and Experimental Setup

The proposed uSDN has been thoroughly evaluated with two widely used benchmark datasets, CAVE [164] and Harvard [22]. The CAVE dataset consists of 32 HR HSI images and each of which has a dimension of 512×512 with 31 spectral bands. These spectral images are taken within the wavelength range $400 \sim 700$nm with an interval of 10 nm. The Harvard dataset includes 50 HR HSI images with both indoor and outdoor scenes. The dimension of the images in this dataset is 1392×1040, with 31 bands taken at an interval of 10nm within the wavelength range of $420 \sim 720$nm. Note that for this dataset, the top left corner of size $1024 \times 1024 \times 31$ is cropped as the HR HSI.

For the two benchmark datasets, the LR HSI $\mathbf{Y}_h$ is obtained by averaging the HR HSI over 32×32 disjoint blocks. The HR MSI images with 3 bands are generated by multiplying the HR HSI with the given spectral response matrix $\mathcal{R}$ of Nikon D700. All the images are normalized between 0 and 1. Note that the CAVE dataset is in general considered a more challenging set than Harvard since images in Harvard usually contain more smooth reflections; and since the images have higher spatial resolution, pixels within close vicinity usually have similar spectral reflectance. Hence, even the images are down-sampled by the 32×32 kernel, most spectral information is still preserved in the LR HSI.

The results of the proposed method on individual images are compared with seven state-of-the-art methods, *i.e.*, CS based [3], MRA based [2], CNMF [166], Bayesian Sparse (BS) [156], HySure [135], Lanaras's 15 (CSU) [87], and Akhtar's 15 (BSR) [4], that belong to different categories of approaches described in Sec. 1. These methods also reported the best performance [100, 4, 87], with the original code made available by the authors. We also directly list results [5] from Akhtar's 16 (HBPG) since the code is not available. The average results on the complete dataset is also reported to evaluate the robustness of the proposed method.

Table 4.2: Benchmarked results in terms of RMSE.

Methods	CAVE					Harvard	
	balloon	CD	cloth	photo	spool	img1	imgb5
CS	25.4	19.4	22.0	18.2	25.8	16.7	17.8
MRA	12.5	14.2	15.4	4.8	11.3	4.7	8.9
BS	14.2	15.3	17.6	11.3	15.2	10.9	14.7
Hysure	14.9	20.3	14.8	4.6	12.5	4.4	5.4
BSR	2.6	7.9	4.3	2.1	6.2	2.3	2.5
CNMF	9.0	11.9	10.1	5.2	12.2	3.2	4.5
CSU	13.3	10	6.7	3.1	7.9	2.2	2.6
uSDN	**1.8**	**4.8**	**3.7**	**2.0**	**5.3**	**2.0**	**0.7**
HBPG	1.9	5.3	3.7	–	–	2.2	0.8

For quantitative comparison, the root mean squared error (RMSE) and spectral angle mapper (SAM) are applied to evaluate the reconstruction error and the amount of spectral distortion, respectively.

4.4.2 Experimental Results

Tables 4.2 and 4.3 show the experimental results of 7 groups of images from the CAVE and Harvard datasets, which are commonly benchmarked by existing literature [77, 4, 5]. We observe that traditional CS-based and MRA-based methods suffer from spectral distortion, thus could not achieve competitive performance. The Bayesian based approach, BS [156], fails due to the fact that it assumes the representation $\mathbf{S}_m$ follows a Gaussian distribution, which is not always true. However, the Bayesian non-parametric based method BSR [4] outperforms BS because it estimates the spectra through non-parametric learning. The matrix-based approaches,CNMF [166] and CSU [87], are not as competitive on the CAVE dataset due to their predefined down-sampling function, although they perform much better on the Harvard dataset. We also observe that some methods like Hysure can achieve better RMSE, but worse SAM scores, that is because they cannot preserve the spectral information properly which has caused large spectral distortion. Based on the experiments, the proposed uSDN powered by the unique sparse Dirichlet-net outperforms all of the other approaches in terms of both RMSE and SAM, and it is quite stable for different types of input images.

Table 4.3: Benchmarked results in terms of SAM.

Methods	CAVE					Harvard	
	balloon	CD	cloth	photo	spool	img1	imgb5
CS	19	17	17	82	48	15	14
MRA	12	9	11	14	15	13	15
BS	11	16	10	18	24	17	18
Hysure	18	24	18	19	38	18	19
BSR	11.9	17.9	6	14	16	1.9	3.4
CNMF	10	9	7	11	20	10	13
CSU	8.9	25	12.6	10	17	1.8	2.8
uSDN	**4.7**	**10**	**4.8**	**5.4**	**13**	**1.6**	**1.7**
HBPG	7.6	10.6	5.0	–	–	2.5	2.1

Table 4.4: The average RMSE and SAM scores over complete benchmarked datasets.

Methods	CAVE		Harvard	
	RMSE	SAM	RMSE	SAM
CSU[87]	9.96	15.63	3.37	5.35
BSR[4]	5.29	13.63	2.61	4.46
uSDN	**4.09**	**6.95**	**1.78**	**4.05**

To further demonstrate the robustness of the proposed uSDN, we report the mean of RMSE and SAM over the complete CAVE and Harvard dataset in Table 3. We only list the performance of matrix factorization based CSU and Bayesian based BSR, since they demonstrated better performance as shown in Tables 1 and 2. We observe that since BSR estimates the representations separately from the spectral bases, although it can achieve good RMSE scores, its SAM scores are not promising. While CSU relates the representations with a predefined down-sampling function, and thus achieves better results on the Harvard dataset, it generates worse results on the CAVE dataset. Both methods may cause spectral distortion in different scenarios. The proposed approach consistently outperforms the other methods in terms of both RMSE and SAM as reported in Table 4.4. We also make two further observations. First, since the Harvard dataset is less challenging than the CAVE dataset, the improvement on the former is not as apparent as that on the latter. This, on the other hand, demonstrates that the proposed uSDN can handle challenging scenarios much better than state-of-the-art. Second, the proposed approach is very effective

in preserving the spectral signature of the reconstructed HR HSI, showing much improved performance especially on SAM on CAVE. The main reason that contributes to the success of the proposed approach is that it relates the representations $\mathbf{S}_h$ and $\mathbf{S}_m$ with statistics and angular difference, *i.e.*, both representations are encouraged to follow a Dirichlet distribution, and their angular difference is enforced to be small. In this way, both the reconstruction error and spectral distortion are effectively reduced.

To visualize the results, we show the reconstructed samples from CAVE and Harvard taken at wavelengths 460, 540, and 670 nm in Fig. 4.5. The first through fourth columns show the LR images, reconstructed images from our method, ground truth images, and the absolute difference between the images at the second and third columns, respectively. We also compare the proposed method with CSU and BSR on the challenge dataset CAVE and show the results in Fig. 4.6. The effectiveness of the proposed method can be readily observed from the difference images, where the proposed approach is able to preserve both the spectral and spatial information.

Convergence Study: During the optimization, both networks converges smoothly as shown in Fig. 4.8. The MSI network has a little bit fluctuation which is caused by the angular difference which is minimized every 10 iterations between the representations of two modalities.

Effect of Free Parameters: There are two free parameters in the algorithm design, i.e., μ for the decoder weight loss and λ for the sparsity control, as shown in Eq. (4.10). We keep $\mu = 1e^{-6}$ during the experiments. To evaluate the effect of sparse parameter, we take the pompom example from the CAVE dataset. Fig. 4.7 shows how RMSE is decreasing when we increase λ from 2×10^{-7} to 1×10^{-6}. We set $\lambda = 1 \times 10^{-6}$ in the experiments.

Visualizing $\mathbf{S}_m$ and $\mathbf{\Phi}_h$: The proposed structure is based on the assumption that the LR HSI, HR MSI, and HR HSI can be formulated as a linear combination of their corresponding spectral bases. Here, we would like to provide visualization results of the spatial representation, $\mathbf{S}_m$, its sparsity property, and the spectral bases, $\mathbf{\Phi}_h$. We use the pompom image from the CAVE dataset as the testing image to generate all the visualization. In order to visually see if the linear combination assumption is valid or not, we project the estimated bases, $\mathbf{\Phi}_m$ into a 3D space using singular value decomposition. In Fig. 4.10, we

observe that the learned bases from CSU is a little bit far away from the data, while the bases from BSR cluster with each other and do not cover all the data. The bases from our method circumscribe the entire data, indicating a more effective representation of the data. Fig. 4.11 shows $\mathbf{S}_m$, which is of the same size as the HR MSI. The pixel value in this image indicates the mixing coefficient (between 0 and 1). We also study if $\mathbf{S}_m$ is indeed sparse or not. The histogram of the learned representations $\mathbf{S}_m$ is shown in Fig. 4.9, where the sparsity is clearly evident.

Comparison to Autoencoder: If we replace the encoder part with traditional autoencoder, the results are shown in Fig. 4.12. It demonstrates that traditional autoencoder could not extract meaningful information from LR HSI due to the lack of necessary constraints enforced on the hidden layer, especially when the dimension of the representations is higher than that of the input, *i.e.*, $l < c$, and the representation layer $\mathbf{S}_m$ tends to be deactivated. On the other hand, the proposed approach encourages the representations to follow a sparse Dirichlet distribution that is able to learn more meaningful representations while minimizing the spectral distortion. Therefore, the difference image between the predicted image and the ground truth has less outliers and spectral distortion as compared to that of the autoencoder.
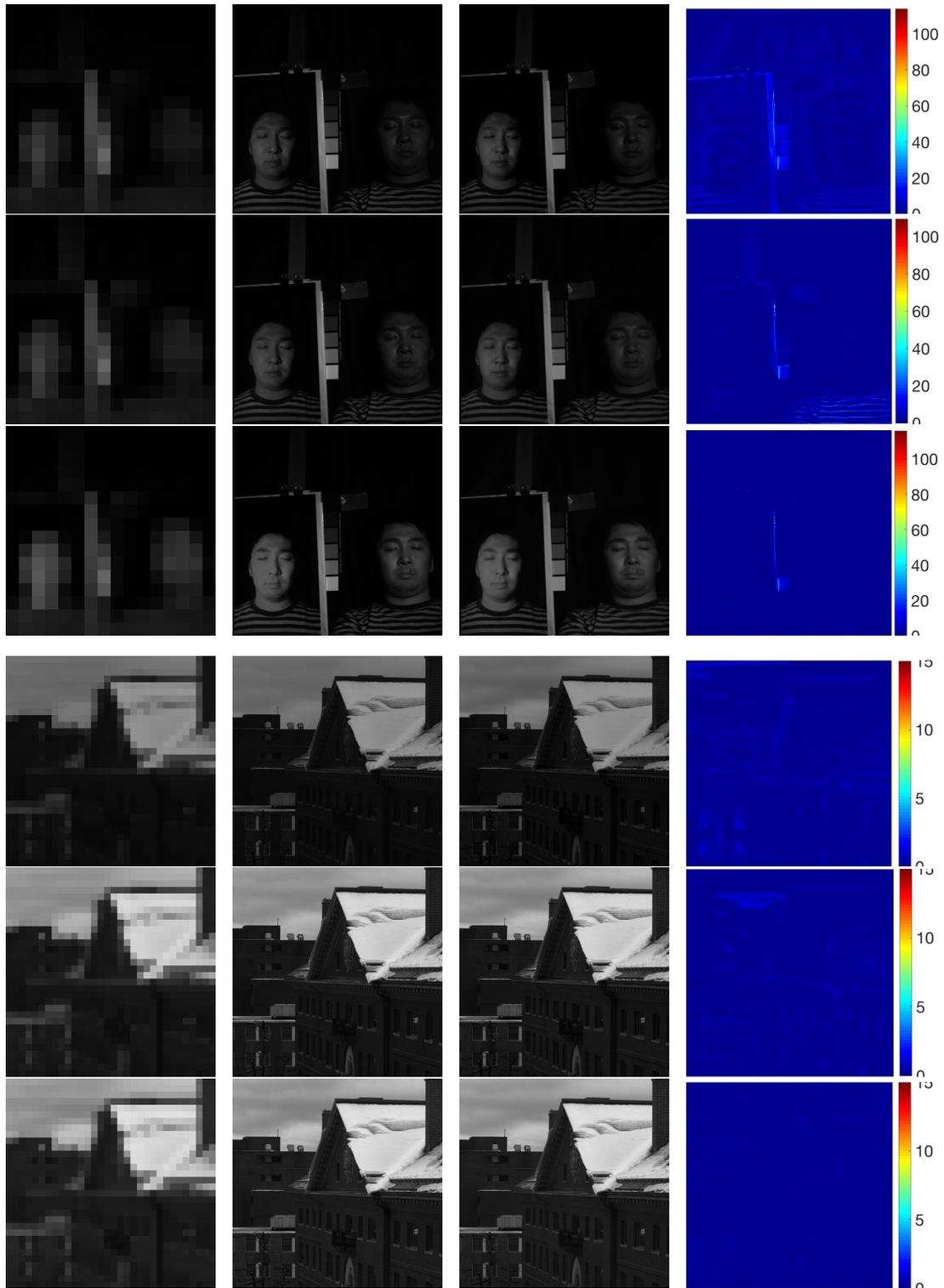
Figure 4.5: Reconstructed images from the CAVE (top) and Harvard dataset (bottom) at wavelength 460, 540 and 620 nm. First column: LR images (16 × 16). Second: estimated images (512 × 512). Third: ground truth images. Fourth: absolute difference.
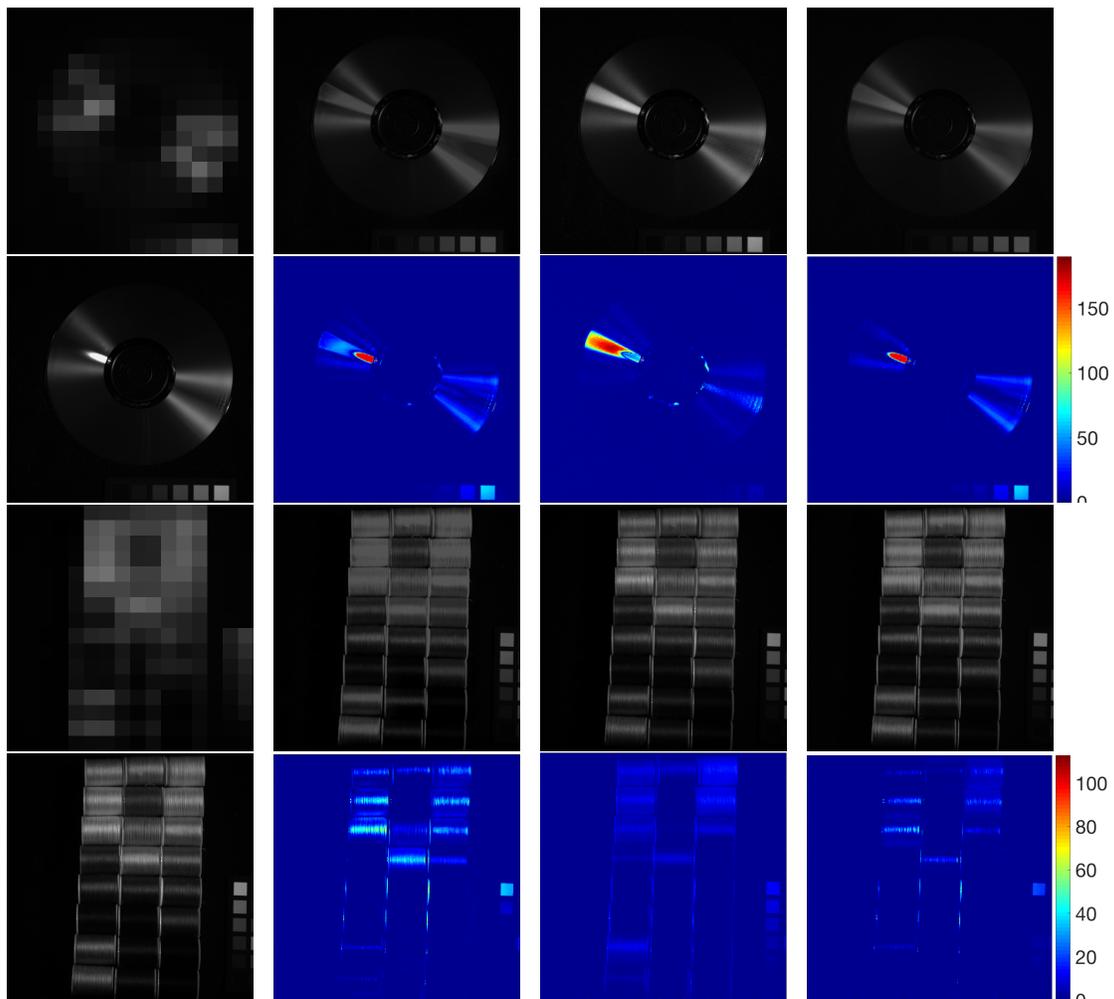
Figure 4.6: Reconstructed images of two examples (top two rows and bottom two rows) from the CAVE dataset at wavelength 670 nm. The first column shows the LR image (top) and the ground truth image (bottom). The second, third and fourth columns are the reconstructed results (top) and the absolute difference (bottom) from CSU, BSR and uSDN, respectively.
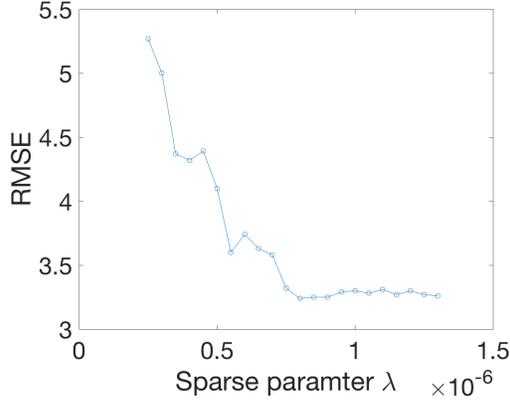
Figure 4.7: The RMSE curve.
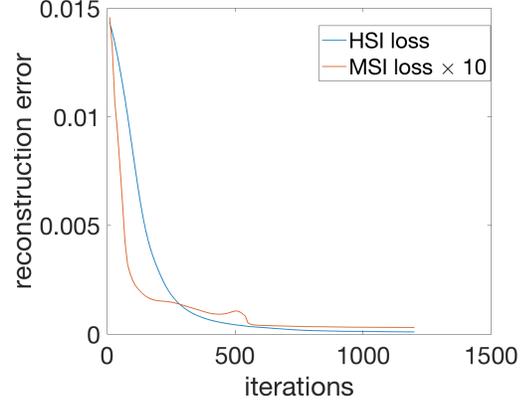
Figure 4.8: Learning curves.
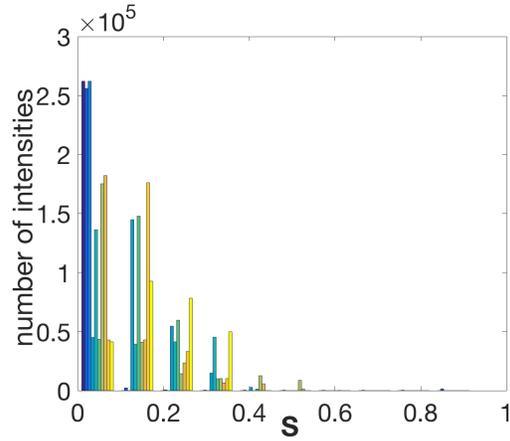
Figure 4.9: Histogram of $\mathbf{S}_m$.
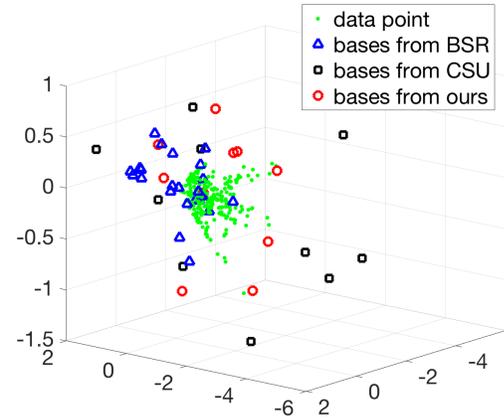
Figure 4.10: Spectral basis.
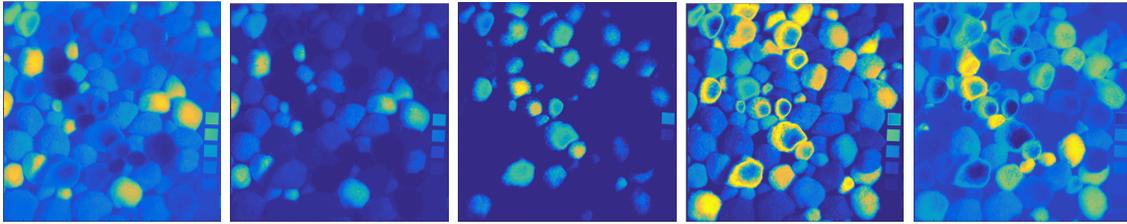
Figure 4.11: Learned representations from the uSDN.
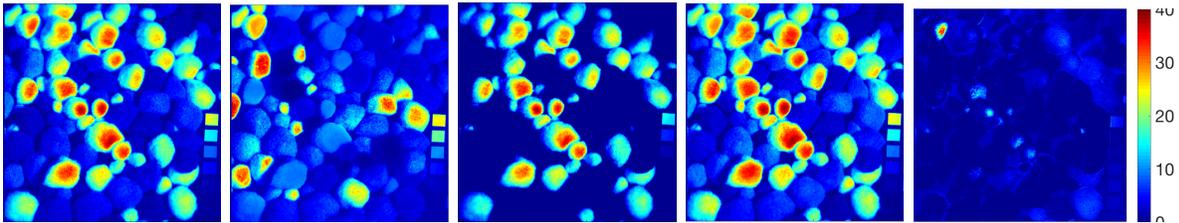
Figure 4.12: Reconstructed results from traditional autoencoder and Dirichlet-Net. From left to right: ground truth image, reconstructed image by the autoencoder, absolute difference, reconstructed image by uSDN and absolute difference.

Chapter 5

Anomaly Detection

5.1 Introduction

Anomaly detection is an unsupervised target detection problem which aims to identify unusual components in the data. It has attracted a lot of attention in remote sensing and has been successfully applied in many applications such as environmental monitoring [144, 17, 16], mineral reconnaissance [18, 105], and national security [141], among others. Since hyperspectral images (HSI) can provide rich spectral characteristics of different materials, which potentially increase the probability of separating anomaly components from background [88, 23, 44, 170], HSI-based anomaly detection has been intensively studied recently.

We can model anomaly detection as an unsupervised binary classification problem between the background class and the anomaly class. The challenge of this problem is that there is no prior knowledge of the anomalies or the background. The background could have complex textures which will increase the difficulty of detection. Also due to the limitations of HSI acquisition devices, HSI data are usually corrupted by noise. Fortunately, there are two important characteristics that could be used to distinguish anomalies from their background, i.e., compared to other objects, the anomalies are 1) sparse, and 2) possess distinctive features as compared to their surrounding background.

Recently, low-rank based methods have also drawn much attention. The basic idea is that, since the anomaly is sparse, the background matrix should have low rank property
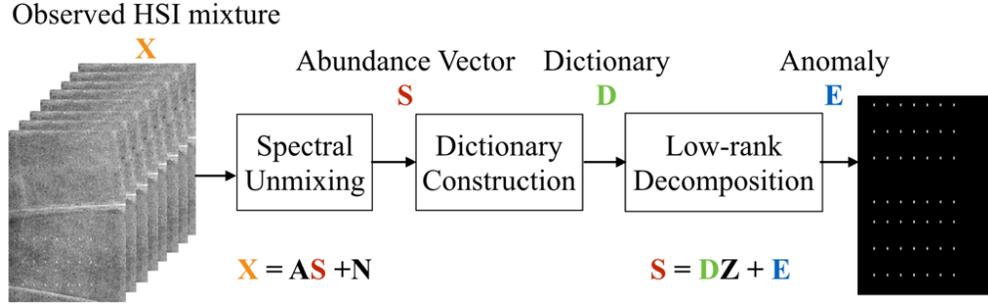
Figure 5.1: Flowchart of the proposed algorithm.

after removing the anomalies [173, 97, 154]. Current approaches [98, 99] perform well in detecting anomalies, but often with high false alarm rate. To overcome the drawbacks of existing approaches, we propose an anomaly detection algorithm based on spectral unmixing and low-rank decomposition. We refer to the proposed anomaly detection algorithm as abundance- and dictionary-based low-rank decomposition, or ADLR.

5.2 Abundance- and Dictionary-based Low-Rank Decomposition (ADLR)

As illustrated in Fig. 5.1, there are three featured components in the proposed ADLR algorithm that contribute to the performance gain, namely, 1) using the abundance vectors derived from spectral unmixing instead of the raw data for anomaly detection, 2) using mean-shift clustering on abundance vectors to construct a dictionary describing both the background and the anomaly, and 3) using a dictionary-based low-rank decomposition to extract the anomaly.

5.2.1 Abundance Extraction through Spectral Unmixing

In HSI data, the measured spectrum at one pixel is usually a mixture of spectra of several constituent materials (known as endmembers). Most mixing processes can be well approximated using a linear mixing model shown below,

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \tag{5.1}$$

66

where $\mathbf{X} \in \mathbb{R}^{l \times n}$ is the observation matrix with each column being the measured spectrum of one pixel, $\mathbf{A} \in \mathbb{R}^{l \times c}$ is the endmember source matrix consisting of c endmembers, $\mathbf{S} \in \mathbb{R}^{c \times n}$ is the abundance matrix, and $\mathbf{N}$ represents the noise.

The unmixing process would, on one hand, try to remove noise $\mathbf{N}$; but more importantly, the abundance matrix extracted serves as a better input to anomaly detection. Each column of the abundance matrix, i.e., the abundance vector, shows how the materials are mixed to form the spectrum of a single pixel. Each row of the abundance matrix, i.e., the abundance map, corresponds to the mixing coefficients of each of the endmembers at the corresponding pixel location of the hyperspectral image. The abundance vectors lie in a single subspace, and in this space, the mixing coefficients for the anomalies present more distinctive patterns as compared to the coefficients of the background. We illustrate this effect in Fig. 5.2 where 20 abundance maps (corresponding to the 20 endmembers) of one of the testing images (Fig. 5.9) are displayed. The anomalies in this test image are the four aluminum panels made with different colors (See Sec. 5.3.1 for more details). The abundance maps of these anomaly endmembers are highlighted in Fig. 5.2. From the figure, we observe that the abundance maps corresponding to the anomaly endmembers show distinctive patterns that can be elaborated with three unique characteristics. First, the foreground pixels (i.e., from the anomaly endmembers) often cluster around a small region. Second, the clustered foreground pixels display high contrast to the background pixels. Third, the foreground pixels usually have a sparse presentation as compared to the noise pixels that usually distribute across the entire abundance map. We will show that recognizing these patterns in the space of abundance vectors is more effective than doing it on the raw HSI images.

Since there is no prior knowledge about the data, unsupervised unmixing is needed in the proposed algorithm. Among the many efficient unsupervised unmixing techniques, such as the ones reported in [109, 110, 15, 127, 148, 24, 73, 17, 170, 58, 168, 122, 106, 95], in this study, we employ the classic method of minimum-volume-constrained nonnegative matrix factorization (MVC-NMF) [109] due to its effectiveness in handling highly mixed HSI data. Note that we also evaluated other more recent unmixing methods, such as [122, 95]; but on
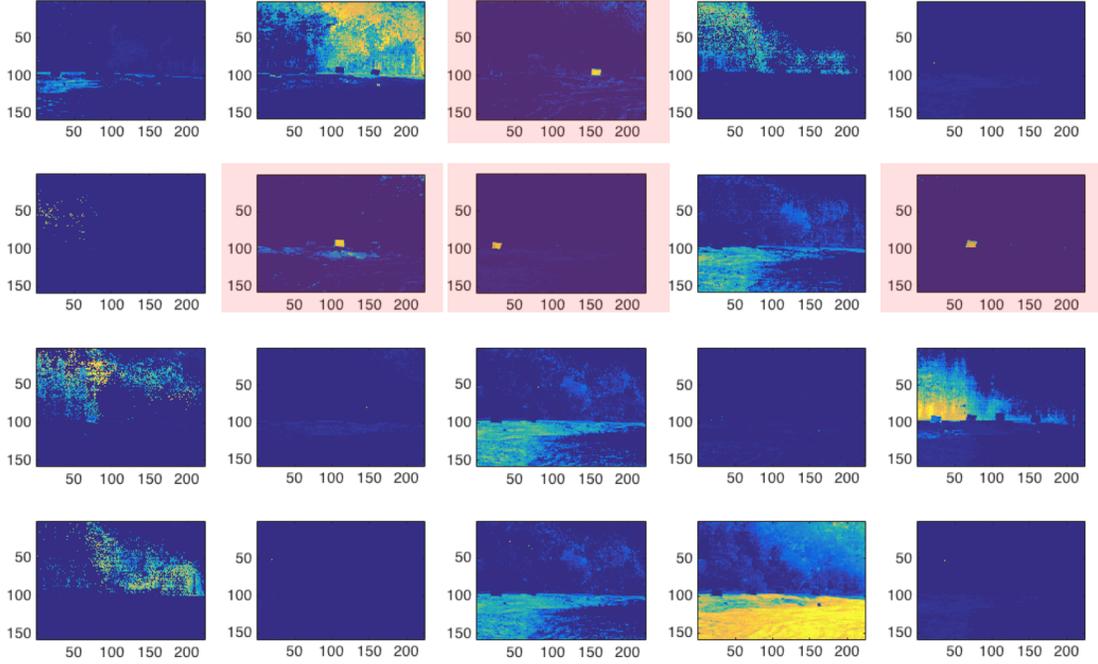
Figure 5.2: Color-coded abundance map of the second test image as shown in Fig. 5.9. The coefficients of the anomalies (highlighted with red in the figure) behave differently from those of the background.

real data, the anomalies are more distinctive in the abundance map estimated by the MVC-NMF. Thus it presents better anomaly detection results than the approaches [122, 95]. See more discussions in Sec. 5.3.4.

MVC-NMF belongs to the family of convex geometry-based unmixing, where the optimal solution is either defined as the one that circumscribes the data cloud and, at the same time, has the minimum volume, or defined as the one that inscribes the data cloud with the maximum volume. Combining the goal of minimum approximation error with the volume constraint, MVC-NMF is formulated as

$$
\min f(\mathbf{A}, \mathbf{S}) = \frac{1}{2}\|\mathbf{X} - \mathbf{A}\mathbf{S}\| + \rho J(\mathbf{A})
$$
$$
\text{s.t.} \quad \mathbf{A} \geq 0 \quad \mathbf{S} \geq 0 \quad \mathbf{1}_c^{\mathrm{T}}\mathbf{S} = \mathbf{1}_N^{\mathrm{T}}
$$
(5.2)

where $\mathbf{1}_c(\mathbf{1}_N)$ is a $c(N)$-dimensional column vector of all 1s, and $J(\mathbf{A})$ is the penalty function, calculating the simplex volume determined by the estimated endmembers. In the objective function, the first term serves as an external force to increase the volume of the simplex

so that it is large enough to comprise all the data point, and the second term serves as the internal force to decrease the volume of the simplex so that we could obtain the simplex with the minimum volume. The parameter ρ is the trade-off between both forces which balance the reconstruction accuracy and the volume constraint [109].

5.2.2 Dictionary-based Low-Rank Decomposition

In the abundance map acquired from the unmixing procedure, we already observed the different behaviors of anomalies, with a highly-contrast, clustered, and sparse presence, as compared to the background including noise. The most challenging issue in anomaly detection in HSI is the existence of noise in the anomaly map. To tackle this problem, we propose a dictionary based low-rank approach to take advantage of the strong representative and discriminative power of the dictionary and we also impose the low-rank constraint on the coefficient of the dictionary, but a sparse constraint on the residue matrix that only contains the anomaly. In the following, we elaborate on how the dictionary is constructed and how the dictionary-based low-rank decomposition problem is solved.

Dictionary Construction through Mean-Shift Clustering

There are many ways to construct a dictionary. The most common one is through dictionary learning [1, 74]. However, in the case of the studied problem, we do not have prior knowledge of either the anomaly or the background, and there is are training data provided, thus the method of dictionary learning is not applicable here. Xu *et al.* [162] proposed to construct the dictionary based on k-means clustering [61] on the original HSI data. In that algorithm, the authors defined the cluster number k to be larger than the true number of ground materials (i.e., the endmembers). Then the RX detector is applied to determine the pre-predictive value of each pixel, where a small pre-predictive value indicates a background pixel. However, making the assumption of known k is often unrealistic, especially for some HSI data with complex background structures.

In the proposed ADLR, the so-called mean-shift clustering method [35] is used on the abundance vectors to automatically derive a set of clusters. The mean-shift algorithm was

originally proposed for mode detection [51]. It considers the set of input points as samples drawn from an empirical probability density function. The algorithm iteratively shifts each data point to the "mode" (or local maximum) of the probability density function until convergence.

Unlike the k-means algorithm, mean shift does not assume prior knowledge of the number of clusters, due to the nonparametric nature of the algorithm. The number of modes automatically indicates the number of clusters. In addition, since mean shift clustering is based on density estimation, it can work on arbitrarily shaped clusters, e.g., the data do not need to follow a fixed distribution, like Gaussian. The key parameter in the mean-shift algorithm is the so-called *window size* or *bandwidth* that specifies the radius or size of the window/kernel function, where the mean value is calculated within each window. The window size determines, indirectly, the distance between clusters. Hence, instead of having to pre-define the number of clusters as in the k-means algorithm, the mean-shift algorithm automatically yields the clustering results based on the specified bandwidth [137].
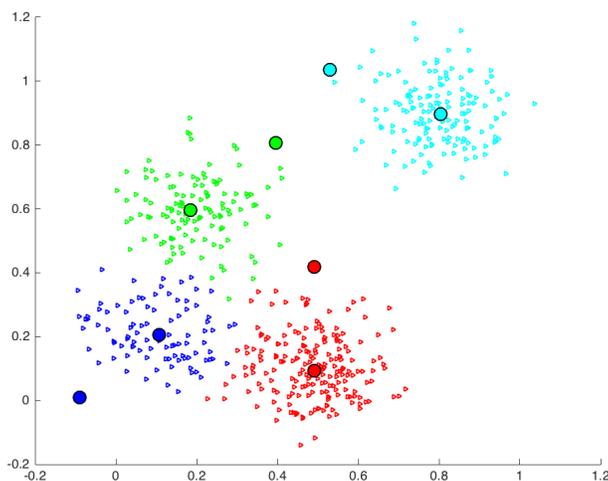
Figure 5.3: Illustration of the inclusion of both the center and edge members of each cluster as dictionary atoms.

Taking the columns of the abundance matrix $\mathbf{S} \in \mathbb{R}^{c \times n}$ as samples, we apply the mean-shift clustering to derive a set of clusters, say k. Normally, we would use the centroid of each cluster as the representative to form one dictionary atom. However, the centroid alone does not reflect the size or shape of the cluster, which serves as important information to distinguish between background (including noise) and the anomaly. Intuitively, anomaly

clusters are dense with the radii of the clusters being much smaller than those of background. To incorporate this intuition into the construction of dictionary, we choose to include both the centroid and the edge member (i.e., the sample with the longest distance to the centroid) of the cluster as the dictionary atoms in **D**. This is illustrated in Fig. 5.3. The inclusion of the edge member might not reflect all the unique characteristics of anomaly clusters (e.g., the shape), however, it well conveys the size of the cluster.

We apply the mean-shift clustering method on the abundance map of the second test image. With the bandwidth selected as 0.2, 41 clusters are generated. The four clusters corresponding to the four anomalies in Fig. 5.2 are illustrated in Fig. 5.4a where different colors represent the different anomaly clusters. The other clusters belong to the background or noisy pixels. It is clear that the isolated pixels in the upper-left corner of the abundance image are grouped into a same cluster. For visualization purpose, we project the clustered data onto a three-dimensional space using the singular value decomposition (SVD) method and the result is shown in Fig. 5.4b. We observe that, in general, the radii of the anomaly clusters are much smaller than those of the background and noise clusters.

(a)   (b)
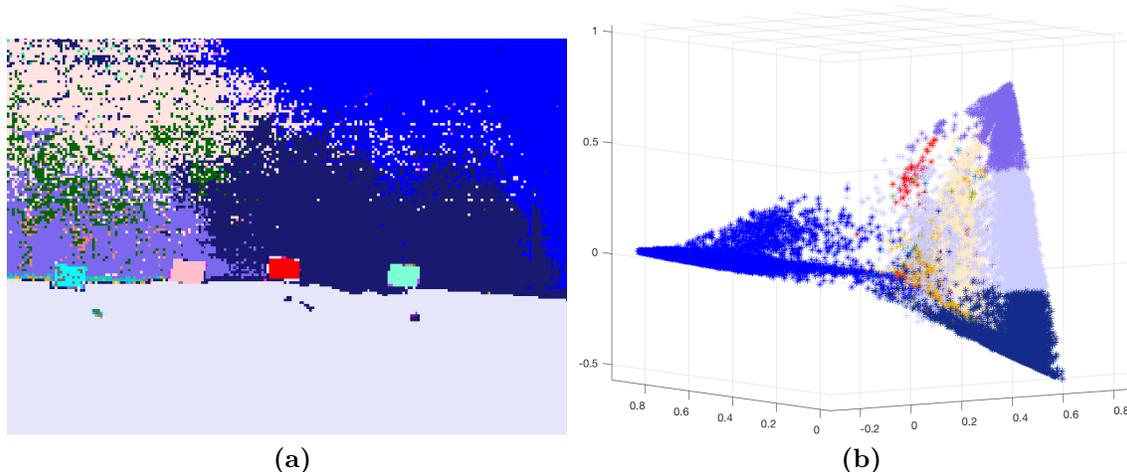
Figure 5.4: Cluster results of the real example from Air Force data shown in Fig. 5.9a. (a) Mean-shift clustering results of the abundance vectors of the second test image. Note that 41 clusters are automatically generated given the bandwidth of 0.2. (b) Mean-shift clustering results of the abundance vectors of (a) after SVD projection

Since there are 41 clusters generated using the second test image, after SVD projection, the cluster boundary is not very clear. In the following, we also provide a toy example to

further demonstrate the different characteristics of the anomaly clusters and the background and noise clusters. In this example, we use the linear mixing of six endmembers, shown in Fig. 5.5c, to generate the synthetic image. Fig. 5.5a illustrates one band of the synthetic data. The background of this data is constructed using four endmembers colored with blue, red, green and cyan, disrupted by adding Gaussian white noise colored with magenta. The standard deviation of the noise is set as 0.047, and the values are limited between 0 and 0.2. The ground truth anomalies are presented in Fig. 5.5b. We again use the SVD method to project the data onto a three-dimensional space and the result is as shown in Fig. 5.5d. Note that, to better illustrate the results, we need the number of clusters to be small, thus we use a relatively larger bandwidth in the mean-shift clustering method. Six clusters are generated given a bandwidth of 0.6. The filled circles with different colors represent the centers and edges of the clusters. The left four clusters belong to the background, the middle cluster belongs to the background with isolated noise and the right cluster marked with yellow color belongs to the anomaly. Note that although there are nine anomalies, after applying SVD, they are projected onto a single yellow dot.

We can observe that, the radius of the anomaly is relatively small, but the radii of the clusters belonging to the background and noise are much larger. In addition, although noise is scattered across the entire image, after the clustering process, they are assigned to the same cluster, thus reducing the probability of identifying isolated pixels as sparse anomalies.

Low-Rank Decomposition

Upon having a better representation scheme of the anomaly through abundance extraction and dictionary construction, we are now ready to apply the low-rank decomposition technique to separate anomalies from the background. There are many ways to address this problem, including, for example, RPCA [98], GoDec [180] and LRR [99]. Table 5.1 shows the different formulations of these three models. The model of RPCA [98] is to decompose a matrix into a low-rank matrix and a sparse matrix via the SVD operation. But the drawback of this model is that isolated pixels are easily identified as anomalies $\mathbf{E}$. Compared to RPCA, the GoDec algorithm also considers the additive noise of the HSI data. But in real HSI data, anomaly $\mathbf{E}$ and noise $\mathbf{N}$ many times are mixed with each other. Even with the unmixing
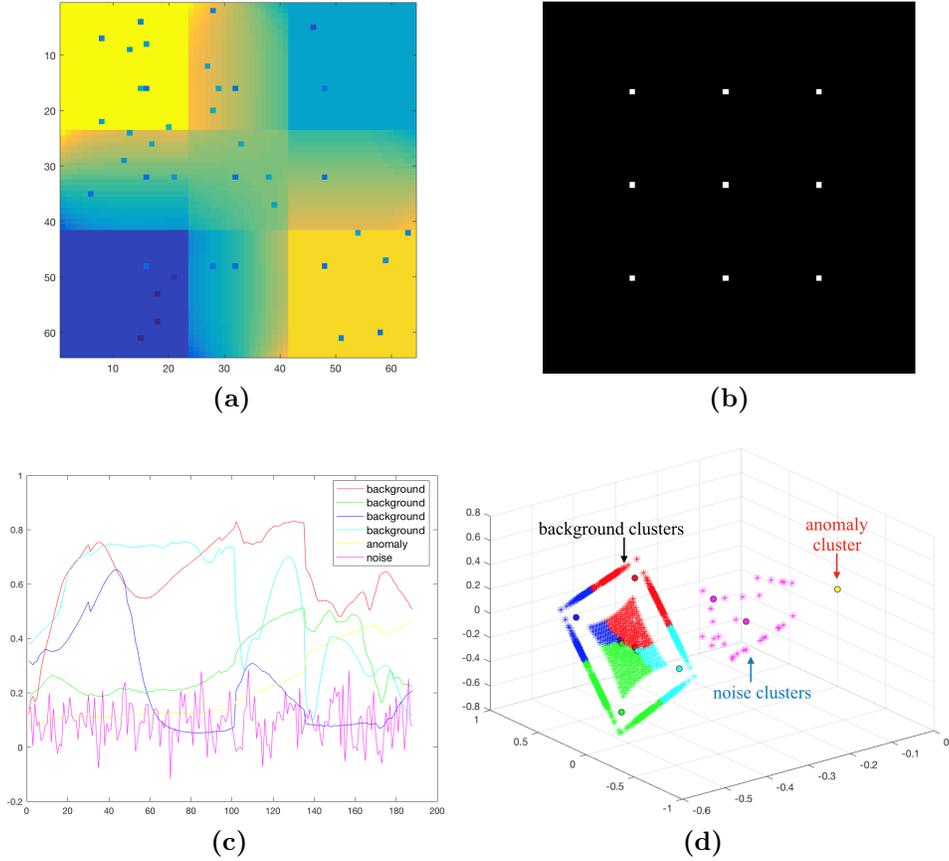
Figure 5.5: Toy example. (a) Synthetic data with noise and anomaly. (b) Ground truth anomaly of (a). (c) Endmembers that generated data (a). (d) Mean-shift clustering results of the abundance vectors of (a) after SVD projection.

algorithm, the noise can only be removed to some extent. LRR [99] is an abbreviation of low-rank representation, which is able to extract sparse outliers in the data with low-rank and l_{21} regularizer. It uses the observed data as the dictionary and works well on data lying in multiple subspaces.

Our model is inspired by LRR [99], where the abundance matrix $\mathbf{S}$ can be further decomposed into two parts:

$$\mathbf{S} = \mathbf{DZ} + \mathbf{E} \tag{5.3}$$

where $\mathbf{D} = \{d_1, d_2, ...d_{2k}\} \in \mathbf{R}^{c \times 2k}$ is the dictionary and $\mathbf{Z} = \{z_1, z_2, ...z_n\} \in \mathbf{R}^{2k \times n}$ is the coefficient for $\mathbf{S}$ with respect to $\mathbf{D}$. $\mathbf{DZ}$ together represents the background. And $\mathbf{E}$ is the anomaly component.

Table 5.1: Low-rank Decomposition Method.

Method	Models
RPCA [98]	$\mathbf{S} = \mathbf{L} + \mathbf{E}$
GoDec [180]	$\mathbf{X} = \mathbf{L} + \mathbf{E} + \mathbf{N}$
LRR [99]	$\mathbf{X} = \mathbf{XZ} + \mathbf{E}$

To solve the problem in Eq. (5.3), we need to introduce some prior knowledge for regularization purpose. Based on the aforementioned analysis, anomalies only occupy a small fraction of HSI, which means $\mathbf{E}$ should be sparse. Since in HSI data, background is highly correlated and can be represented by a dictionary, the coefficient $\mathbf{Z}$ should have low-rank.

Although LRR [99] can be used to address the problem in Eq. (5.3), it is based on the assumption that the data are drawn from multiple subspaces, so it uses the l_{21} norm to force most of the columns of $\mathbf{Z}$ to be zero. In the proposed method, $\mathbf{S}$ is the abundance vector obtained from the unmixing step which lies in a single subspace. In addition, anomalies in the same image can present different characteristics. Thus it is reasonable that the columns of matrix $\mathbf{E}$ have a few nonzero values. Based on the analysis above, we use the l_1 norm to encourage matrix $\mathbf{E}$ to be sparse. Compared to the l_{21} norm, the l_1 norm allows nonzero values to exist in columns.

It has been proven in [160] that minimizing the nuclear norm of a matrix is a good surrogate for minimizing the rank of a matrix, as long as the matrix $\mathbf{E}$ is sufficiently sparse. Therefore, the proposed objective function can be written as

$$
\min_{\mathbf{Z},\mathbf{E}} \quad \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1
$$

$$
s.t. \quad \mathbf{S} = \mathbf{DZ} + \mathbf{E} \tag{5.4}
$$

where $\|\mathbf{Z}\|_*$ denotes the nuclear norm, i.e., the sum of singular values of $\mathbf{Z}$, and $\|.\|_1$ is the l_1-norm. $\lambda > 0$ is the parameter used to balance the trade-off between the low rank and sparsity.

Optimization

Due to the usage of different regularizers and the different dictionary construction, the optimization process is accordingly different from that in LRR. In order to separate the variables $\mathbf{D}$ and $\mathbf{Z}$ in the optimization problem shown in Eq. (5.4), we follow the approach in [99] and introduce an auxiliary variable J as a split variable to replace Z. Therefore, the objective function becomes

$$
\min_{\mathbf{Z},\mathbf{E},\mathbf{J},\mathbf{Y}_1,\mathbf{Y}_2} \|\mathbf{J}\|_* + \lambda \|\mathbf{E}\|_1 +
$$
$$
< \mathbf{Y}_1^T, \mathbf{S} - \mathbf{DZ} - \mathbf{E} > + < \mathbf{Y}_2^T, \mathbf{Z} - \mathbf{J} > + \qquad (5.5)
$$
$$
\frac{\mu}{2} \left(\|\mathbf{S} - \mathbf{DZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2 \right)
$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrange multipliers, $\mu > 0$ is the penalty parameter. Eq. (5.5) can be solved alternatively by updating only one variable a time while fixing the others. The detailed procedure is described in Algorithm 2.

- **J step:** Fix $\mathbf{Z}$ and $\mathbf{E}$ to update $\mathbf{J}$. The objective function with respect to $\mathbf{J}$ can be written as:

$$
\min_{\mathbf{J}} \|\mathbf{J}\|_* + < \mathbf{Y}_2^T, \mathbf{Z} - \mathbf{J} > + \frac{\mu}{2} \|\mathbf{Z} - \mathbf{J}\|_F^2
$$
$$
= \min_{\mathbf{J}} \frac{1}{\mu} \|\mathbf{J}\|_* + \frac{1}{2} \|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}_2/\mu)\|_F^2 \qquad (5.6)
$$

and Eq. (5.6) can be solved by Lemma 3.2 in [99].

- **Z step:** Fix $\mathbf{J}$ and $\mathbf{E}$ to update $\mathbf{Z}$. The objective function with respect to $\mathbf{Z}$ can be written as:

$$
\min_{\mathbf{Z}} < \mathbf{Y}_1^T, \mathbf{S} - \mathbf{DZ} - \mathbf{E} > + < \mathbf{Y}_2^T, \mathbf{Z} - \mathbf{J} >
$$
$$
+ \frac{\mu}{2} \left(\|\mathbf{S} - \mathbf{DZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2 \right)
$$
$$
\mathbf{Z} = (\mathbf{D}^T \mathbf{D} + \mathbf{I})^{-1} [\frac{1}{\mu} (\mathbf{D}^T \mathbf{Y}_1 - \mathbf{Y}_2) \qquad (5.7)
$$
$$
+ \mathbf{D}^T \mathbf{S} - \mathbf{D}^T \mathbf{E} + \mathbf{J}]
$$

Algorithm 2 Anomaly Detection by the Proposed ADLR

Input: S, D

Output: Z, E

  Initialization: Fix μ and λ according to the given data (See Sec. 5.3.5). $\mathbf{Z} = \mathbf{J} = \mathbf{E} = \mathbf{Y}_1 = \mathbf{Y}_2 = \mathbf{0}$.

  while not converge **do**

  1) Fix $\mathbf{Z}$, $\mathbf{E}$ and update $\mathbf{J}$

    $\mathbf{J} = \arg\min \frac{1}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}_2/\mu)\|_F^2$

  2) Fix $\mathbf{J}$, $\mathbf{E}$ and update $\mathbf{Z}$

    $\mathbf{Z} = (\mathbf{D}^T\mathbf{D} + \mathbf{I})^{-1}[\mathbf{D}^T\mathbf{S} - \mathbf{D}^T\mathbf{E}$

      $+ \mathbf{J} + (\mathbf{D}^T\mathbf{Y}_1 - \mathbf{Y}_2)/\mu]$

  3) Fix $\mathbf{J}$, $\mathbf{Z}$ and update $\mathbf{E}$

    $\mathbf{E} = \arg\min \frac{\lambda}{\mu}\|\mathbf{E}\|_1 + \frac{1}{2}\|\mathbf{E} - (\mathbf{S} - \mathbf{DZ} - \mu^{-1}\mathbf{Y}_1)\|_F^2$

  4) Update $\mathbf{Y_1}$ and $\mathbf{Y_2}$

    $\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathbf{S} - \mathbf{DZ} - \mathbf{E})$

    $\mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{Z} - \mathbf{J})$

  Check the convergence:

    $\|\mathbf{S} - \mathbf{DZ} - \mathbf{E}\|_\infty < \varepsilon$

    and $\|\mathbf{Z} - \mathbf{J}\|_\infty < \varepsilon$

  end

  return $\mathbf{Z}, \mathbf{E}$

- **E step:**   Fix $\mathbf{J}$ and $\mathbf{Z}$ to update $\mathbf{E}$. The objective function with respect to $\mathbf{E}$ can be written as:

$$
\begin{aligned}
\min_{\mathbf{E}} \lambda\|\mathbf{E}\|_1 &+ <\mathbf{Y_1^T}, \mathbf{S} - \mathbf{DZ} - \mathbf{E}> \\
&+ \frac{\mu}{2}\|\mathbf{S} - \mathbf{DZ} - \mathbf{E}\|_F^2 \\
&= \min_{\mathbf{E}} \frac{\lambda}{\mu}\|\mathbf{E}\|_1 + \frac{1}{2}\left\|\mathbf{E} - (\mathbf{S} - \mathbf{DZ} - \mu^{-1}\mathbf{Y_1})\right\|_F^2
\end{aligned}
\tag{5.8}
$$

Eq. (5.8) can be solved through the soft thresholding method as shown in Algorithm 2. After decomposing the abundance map $\mathbf{S} \in \mathbb{R}^{c \times n}$, we can obtain the anomaly matrix $\mathbf{E} \in \mathbb{R}^{c \times n}$. To generate the anomaly image, we sum along each column of $\mathbf{E}$ and reformat into the same spatial dimension of a single image in HSI.

5.3 Experimental Results and Discussions

In this section, the proposed algorithm is evaluated using three hyperspectral image datasets with different characteristics, including one synthetic and two real data. Both qualitative and quantitative comparisons are conducted on these three test images.

5.3.1 Dataset Description

The first test image is a hyperspectral image embedded with synthetic targets. The original image contains 79 bands of dimension 150×103. Forty-nine anomalous dots are implanted based on the linear mixing model [129, 45] defined in Eq. (5.9). The targets are distributed in 7 rows and 7 columns and $\mathbf{t}$ is fractionally implanted with background spectrum $\mathbf{b}$ by varying f from 0.05 to 1 for different rows:

$$\mathbf{z} = f \cdot \mathbf{t} + (1 - f) \cdot \mathbf{b} \tag{5.9}$$

The second test image is provided by Air Force [178, 177] which contains non-uniform background and anomaly targets with different reflections. The image contains 124 bands of dimension 267×342. The anomalies in the image are four aluminum panels with different colors (Black, Green, Tan, and Silver). Note that in this image, the anomaly is bigger than that of the first test image.

The third test image is provided by NASA [8, 36], acquired from the two Mast Cameras (Mastcams) installed on the Mars Science Laboratory (MSL) Curiosity Rover. The two Mastcams can create multispectral images of 12 bands in the visible and near infrared wavelengths. The dimension of the image is 598×670. The anomalies (or objects of interest) in this image are hydrated minerals.

The first two test images have ground truth but the third one does not. The three test images are displayed in Figs. 5.6a, 5.9a, and 5.12a, respectively.

5.3.2 Experimental Design and Performance Metrics

We evaluate the proposed ADLR from two perspectives. First, we compare ADLR with six representative anomaly detection methods, including global RX (GRX) [123], local RX (LRX) [142], and subspace RX (SSRX) [128], which are statistical-based; collaborative representation detection (CRD) [96], which is representation-based; and low-rank and sparse matrix decomposition based Mahalanobis distance method (LSMAD) [173] and the low-rank tensor based method proposed by Li *et al.* [97], which are low rank-based. Note that, since we do not have the codes of LSMAD, we implement it with GoDec [180] and GRX based on the description in LSMAD [173].

In addition, we also perform comprehensive evaluation of the contribution of different components of ADLR. First, we compare ADLR with RPCA that applies low-rank decomposition on the raw image data directly. This comparison would show the benefit of anomaly detection based on the abundance rather than the raw data. Second, we compared ADLR with Wang *et al.* [154], which does apply low-rank decomposition on the abundance, but did not use dictionary to better represent the background. This comparison would show the benefit of dictionary construction and the application of the low-rank constraint on the background coefficient rather than the background directly. Third, we apply LRR with l_{21} constraint on the abundance vectors, whose dictionary is defined by the proposed ADLR. With this experiment, we show that because we have projected the data onto a single subspace through spectral unmixing, the proposed l_1 constraint works better than the l_{21} constraint in the LRR.

Besides qualitative comparison through visual examination of the detection results, for quantitative comparison, detection results are converted to binary images according to different thresholds. Based on these binary images, the receiver operating characteristic (ROC) curves are generated by calculating the detection rate versus false alarm rate. Then the area under the curve (AUC) is calculated according to the ROC curve of each graph. Both detection rate and false alarm are evaluated for comprehensive evaluation. Here, we also compute the highest detection rate and highest false alarm rate. The highest detection rate is defined as the maximum rate among the detection rates from all thresholds. Similarly,

the highest false alarm rate refers to the maximum false alarm rate among the false alarm rates from all thresholds. The larger the highest detection rate, the better the performance of an algorithm is; the lower the highest false alarm rate, the better the performance of an algorithm is.

5.3.3   Anomaly Detection Performance

Synthetic Targets on Real Data

Fig. 5.6 shows the comparison of different algorithms on the synthetic dataset, which is the first test image. We observe that the first row of the simulated targets is always difficult to detect, since the target is fractionally implanted by only 5%. The results from GRX and LSMAD have low contrast between anomalies and background. By visual comparison, the LRX algorithm keeps the best contrast between anomalies and background. This is because LRX applies GRX in a local area, so it performs extremely well on uniformly distributed targets. However, the noise in the background is also mistakenly selected as anomalies, leading to a high false alarm rate. SSRX effectively remove some noise in the background, but not all anomalies are detected. Li's method can identify all the targets, but as well as noise in the background. CRD works effectively on the synthetic dataset which detects most of the anomalies with less noise compared to LRX and Li's method. Because CRD assumes that background pixels can be represented by surrounding background, which is suitable for small anomaly with a uniform background.

To assess the effectiveness of different components in the proposed ADLR, we perform various combinations of the possible components in the proposed algorithm and show the results in Fig. 5.6h-k. Without spectral unmixing, Fig. 5.6h shows the result of applying RPCA method directly on image pixels of Fig. 5.6a. Most of the anomalous points can be detected by the RPCA method. However, just like LRX, the background is also selected as anomalies. With spectral unmixing and RPCA low-rank decomposition [154], the result is shown in Fig. 5.6i. This method can detect the anomalies with little background. But some noisy and isolated pixels are sparse and are confused as anomalies. For example, in the middle of this image, a large area of background is also selected as anomalies. If we

apply LRR with l_{21} constraint on the abundance matrix generated from spectral unmxing, the result is shown in Fig. 5.6j. Some noise is removed, but not all the anomalies are detected because of the fact that l_{21} constraint forces the columns of the anomaly matrix to be zeros. However, in the abundance matrix, not all the elements in the column that represents the anomaly are zeros. Thus, not all the anomalies are extracted during the optimization procedure. As demonstrated in Fig. 5.6k, most of the noise pixels have been removed by the proposed ADLR with spectral unmixing, dictionary construction and low-rank decomposition with l_1 norm, as compared to the result of Fig. 5.6i. The reason is that we only force the coefficients of the background to be low-rank. Since the isolated pixels can be grouped with similar coefficients with respect to the dictionary, they lose the property of sparsity in the image. Therefore, most of the isolated noisy background have been removed from the sparse matrix $\mathbf{E}$. This illustrates that both the dictionary and the l_1 regularizer played an important role in noise removal in the proposed algorithm.

For quantitative comparison, the results of each algorithm are converted to different binary images using thresholds ranging from 0.01 to 1. The false alarm rate and detection rate are calculated based on the given ground truth target and each binary image. Then the ROC curve is generated by plotting the detection rate against the false alarm rate. Note that, for ROC curve, the best possible method should yield a point in the upper left corner (0,1), which indicates that the method can achieve 100% detection accuracy with false alarm rate of 0. As illustrated in the ROC curve given by Fig. 5.7, the LRX method is able to reach the (0,1) point when we adjust the threshold of the image, but it yields high false alarm rate around 1 with other thresholds. Wang's method yields low false alarm rate, but it is not able to detect all the targets. Li's method is capable of detecting all the targets, but like the result from LRX, it also yields high false alarm rate. Compared to the other algorithms, the proposed ADLR identifies all the targets while maintaining a low false alarm rate. Note that, although the proposed method cannot achieve the best performance in the top-left corner as LRX, the benefit of ADLR is that there is no need to adjust the threshold to achieve high detection rate and low false alarm rate because the detected anomalies have high contrast against the background. Given a small threshold like 0.01, we are able to achieve high detection rate and low false alarm rate.

The AUC is also estimated and given in Table 5.2 based on the ROC. The AUC derived from the proposed ADLR is comparable to that of the LRX. Note that the higher the AUC value, the better the detection algorithm is.

Figure 5.8 shows the highest detection rate with the highest false alarm rate of each of the evaluated approaches. We observe that the proposed method has the lowest highest false alarm and a competitive highest detection rate. No matter how we adjust the threshold of the output image, the false alarm rate remains below 0.079.

Table 5.2: AUC for the first and second test images obtained from different algorithms.

Algorithms	Test images #1	Test images #2
GRX	0.5539	0.4941
LRX	**0.9937**	0.5132
SSRX	0.9481	0.5022
Li's	0.9810	0.9853
CRD	0.9456	0.4108
LSMAD	0.8422	0.8857
RPCA only	0.9153	0.3725
Wang's	0.8520	0.9796
Proposed ADRL	0.9820	**0.9867**

Real Data with Nonuniform Background

For the second test data, there are four targets/anomalies made with different colors, and the reflection of the targets are similar to that of the background in several bands. Also, the background of the data is nonuniform which increases the difficulty of detection. Fig. 5.9 shows the detection results of different algorithms. GRX, LRX SSRX and LSMAD fail to detect all the anomalies. This is because RX-based approaches assume that the probability density function of the background can be modeled as the Gaussian distribution, which is not applicable when the background is complex and has different textures. CRD fail in this case as well because the size of the anomaly is big and the background is non-uniform. Both Li's method and Wang's method can successfully detect all the four targets, but they also identify some textures from background as anomalies.

We further evaluate the performance of various combinations of the possible components in the proposed algorithm. Figure 5.9h shows the result by applying RPCA low-rank

decomposition directly on the images. It illustrates that without spectral unmixing, the anomalies and their background have low contrast. As shown later in Fig. 5.10, the detection rate is very low. With spectral unmixing and RPCA low-rank decomposition, four targets are detected successfully, but the third one has low contrast, as shown in Fig. 5.9i. The anomalies identified by LRR given the dictionary defined by the proposed ADLR misses one anomaly because of the l_{21} constraint, as shown in Fig. 5.9j. Compared to other methods, the proposed ADLR including spectral unmixing, dictionary construction and low-rank decomposition detect all the targets with lowest noise level as shown in Fig. 5.9k.

The ROC curve given in Fig. 5.10 and the bar chart in Fig. 5.11 both demonstrate that the proposed approach has a higher detection rate over the other methods. Meanwhile, it keeps the false alarm rate below 0.01 regardless of the threshold used. The AUC of different algorithms is given in Table 5.2, which also indicates that the proposed algorithm achieves the best result among all the algorithms.

Real Data without Ground Truth

The results of the last HSI data are shown in Fig. 5.12. The proposed algorithm successfully detects the hydrated 'anomaly' which mostly appears in the drilled hole and cracks of the soil surface. As shown in Fig. 5.12, the detected target from the proposed ADLR has less background portion as compared to that of Li's method or Wang's method. When we apply the LRR on the abundance vector, with the dictionary defined using ADLR, the detection result has less background portion than that of Wang's method. But the anomaly is not as clear as that of the proposed ADLR. All these four approaches are superior to that of GRX, LRX, SSRX, CRD and LSMAD methods. Note that since we do not have ground truth anomaly for this test image, the ROC curve and AUC are not provided in this experiment.

It is worth mentioning that although LRX can identify anomaly with high contrast on synthetic data, it fails on real data because of its deficiency in handling background that is highly nonuniform and anomalies of arbitrary size.

5.3.4   Effects of Different Unmixing Algorithms

We have addressed the importance of introducing spectral unmixing in anomaly detection and have chosen MVC-NMF to conduct the unmixing procedure. In this experiment, we evaluate the effect of choosing different unmixing algorithms on the detection performance. We compare ADRL using MVC-NMF as well as two other state-of-the-art unmixing algorithms from [122], [95] on the first two test images.

Table 5.3:   AUC for the first synthetic test image obtained with different unmixing algorithms

AUC	MLNMF	R-CoNMF	MVCNMF
Test image 1	0.8588	0.8854	**0.9820**
Test image 2	0.7779	0.6988	**0.9867**

We observe from Fig. 5.13 that all the unmixing approaches achieve relatively clear results with little background. The ROC curves in Fig. 5.14 and AUC in Table 5.3 both demonstrate that all the approaches are able to yield high detection rate with low false alarm rate. In Fig. 5.15, we further observe that the highest false alarm rates are always below 0.4, which is much lower as compared to the traditional algorithms without the unmixing step, e.g., LRX, SSRX, Li's, CRD, LSMAD and RPCA.

Figs. 5.16–5.18 demonstrate the corresponding results on the second test image. We again observe that the highest false alarm rate achieved by these unmixing-based detectors are lower than traditional algorithms that operate on the raw data directly, e.g, RX, Li's, CRD and RPCA.

From experiments conducted on both test images, we see that MVC-NMF constantly presents higher detection rate, lower false alarm rate, and higher AUC values. This is due to the fact that both MLNMF and R-CoNMF use sparsity as one of the constraints. While in synthetic dataset, this might be common; for real hyperspectral data, especially for highly-mixed data, the abundance vector is, in general, not sparse. This is the main reason that MVC-NMF-based detectors can outperform the other two unmixing approaches.

5.3.5 Parameter Analysis

The proposed ADLR has three free parameters, i.e., c, the number of endmembers in the unmixing procedure (note that ρ in Eq. (5.2) is fixed to a default value in this step), bw, the bandwidth in the clustering step, and λ, the trade-off parameter in low-rank decomposition. In addition, in LRX, we set the window size as 11×11. For CRD, we use the default parameters offered by the authors. Table 5.4 shows all the parameters used in the experiments.

Table 5.4: Parameters used in the experiments.

Test image	c	bw	λ	c estimated by HySim
#1	25	0.05	0.01	16
#2	20	0.2	0.02	14
#3	9	0.2	0.1	4

In all the experiments, we first estimated the parameter c with the HySime algorithm [13]. For better unmixing, we set c to be higher than the number of endmembers estimated by HySim [13]. Although increasing the number of endmembers may introduce noise to the data, the noise will be removed in the subsequent dictionary construction and low-rank decomposition steps.

To evaluate how the number of the endmembers affect the results, we fix $bw = 0.2$, $\lambda = 0.1$, and adjust the parameter c. Fig. 5.19 shows the AUC of the first two test images, according to different number of endmembers. Based on the experiments, increasing the numbers of endmembers does not always decrease the AUC value. When the anomalies are separated from the background well, we can achieve high AUC values. For example, for the first synthetic test image, the AUC reaches the highest value when $c = 25$, where the abundance vectors of the anomalies are distinctive from those of the background. But for the second test image, the anomalies are separated well when $c = 20$, therefore, the AUC at $c = 20$ is the highest among all the experiments. For both data, the AUC stays above 0.88 even with different numbers of endmembers. Therefore, the unmixing step effectively increases the anomaly detection rate. But to achieve the best result, it is important to distinguish anomalies in the abundance space. In general, we set c to be around 1.5 to 2 times of the c given by HySime.

We then fix the number of endmembers and adjusted the parameters bw and λ. Based on the experimental study, the parameter bw is evaluated in the range of $[0.05, 1]$, and the parameter λ in the range of $[0.01, 0.1]$. Figs. 5.20 and 5.21 show the change of AUC when we adjusted the free parameters bw and λ for the first two test images, respectively. When bw is small, we obtain more clusters in the dictionary construction step. Then there will be more dictionary atoms for the low-rank decomposition which makes AUC more sensitive to the parameter λ. When $bw \in \{0.8, 0.9, 1\}$, there exists only one cluster of the abundance. So the dictionary atoms is defined by the center and the edge of this cluster. From the AUC curve, we observe that even with a single cluster, we can still yield an AUC of above 0.98 for both datasets no matter how we adjust the parameter λ. But consider the false alarm rate, we should choose a relative small bandwidth bw with a lower λ. Note that the number of clusters varies according to the given data even with the same bandwidth. Therefore, we usually choose the bandwidth that generates less than 100 clusters; in this way, the anomaly detection procedure would be faster.
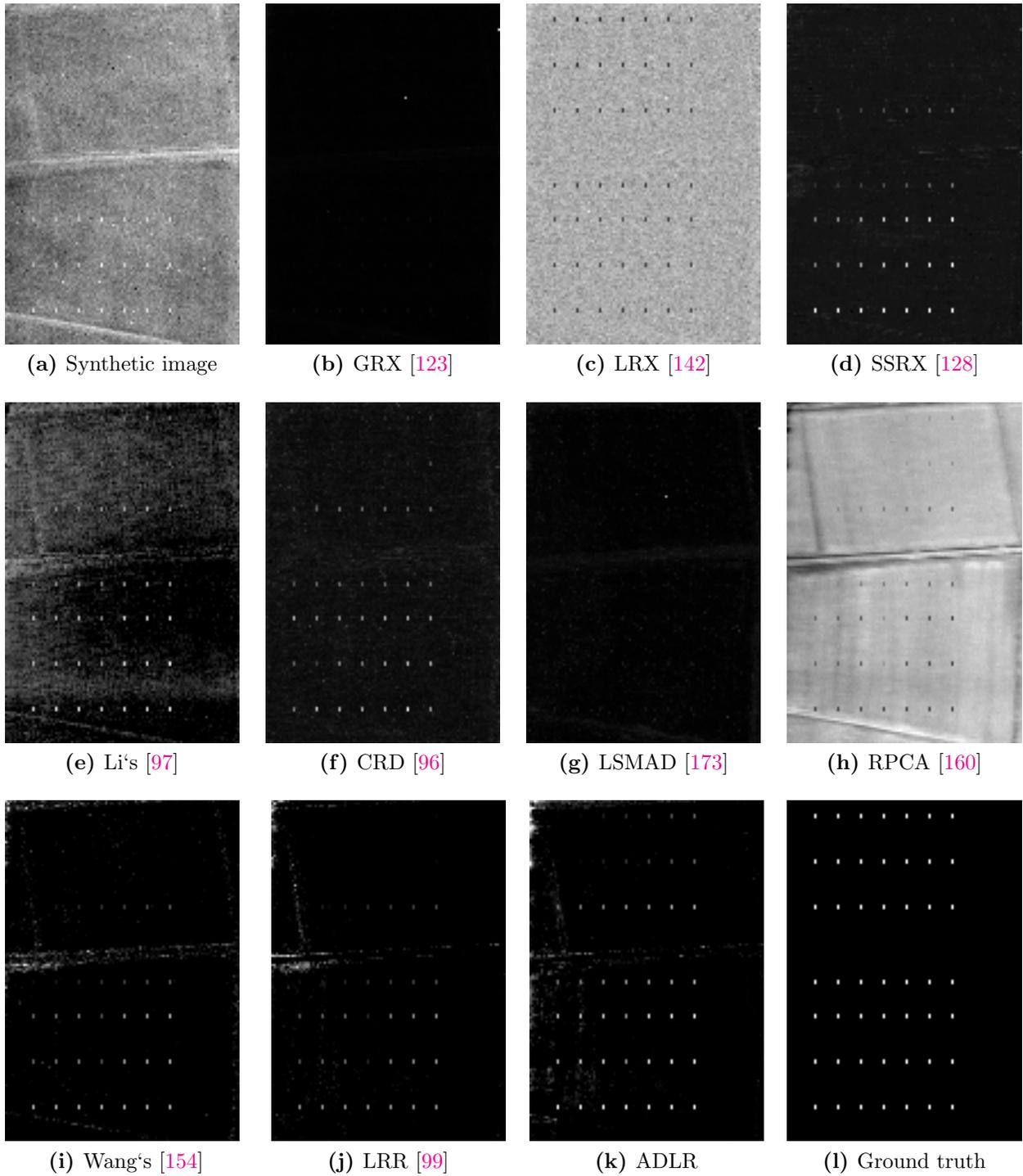
(a) Synthetic image   **(b)** GRX [123]   **(c)** LRX [142]   **(d)** SSRX [128]

(e) Li's [97]   **(f)** CRD [96]   **(g)** LSMAD [173]   **(h)** RPCA [160]

(i) Wang's [154]   **(j)** LRR [99]   **(k)** ADLR   **(l)** Ground truth

Figure 5.6: Detection results using the 9 algorithms on the first test images (synthetic images).
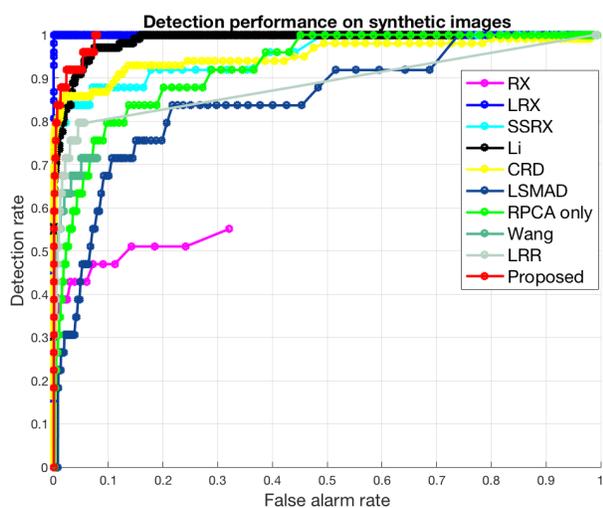
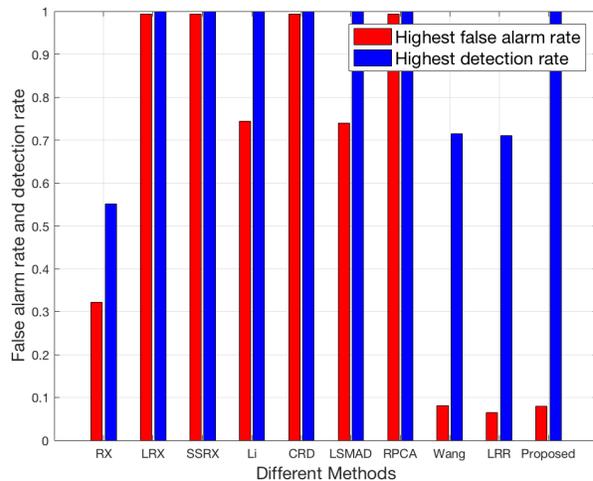**Figure 5.7:** ROC curves of anomaly detection performance with different algorithms on the first test image.
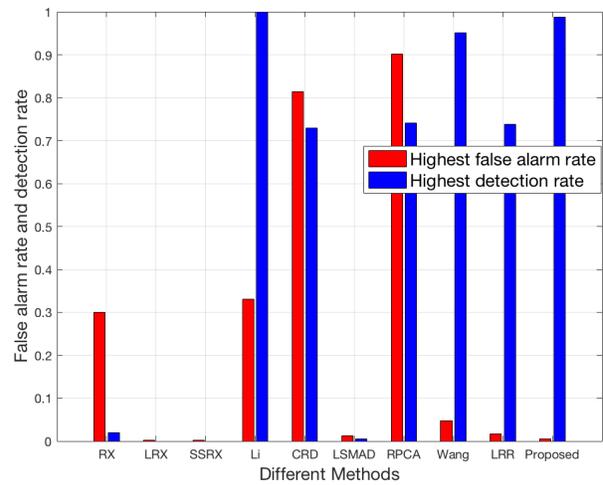

Figure 5.8: Highest detection rates and highest false alarm rates obtained from different algorithms on the first synthetic test image.

(a) Air Force HSI data   **(b)** GRX[123]   **(c)** LRX [142]   **(d)** SSRX [128]

(e) Li's [97]   **(f)** CRD [96]   **(g)** LSMAD [173]   **(h)** RPCA [160]

(i) Wang's [154]   **(j)** LRR [99]   **(k)** ADLR   **(l)** Reference target
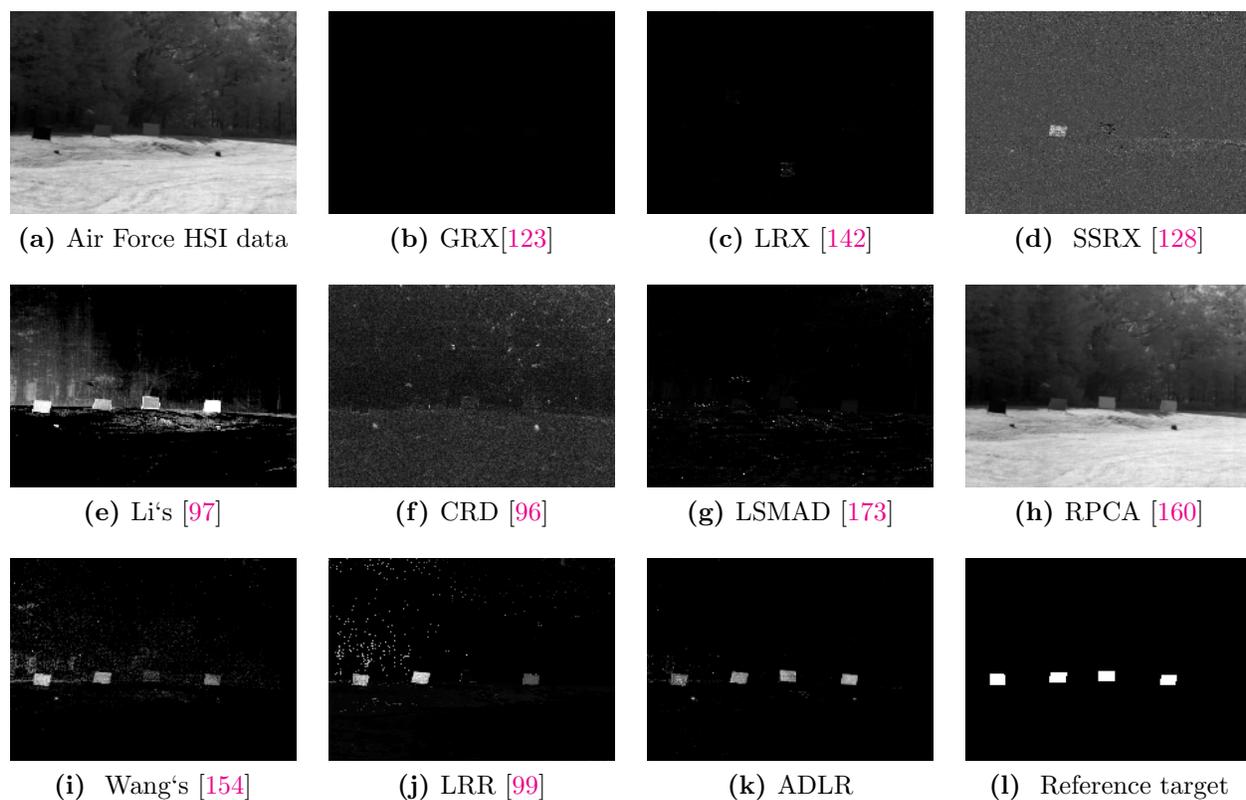
Figure 5.9: Detection results of hyperspectral images from Air Force.
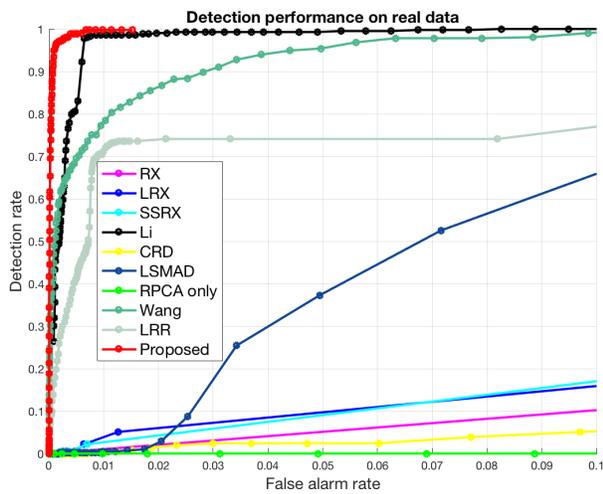
Figure 5.10: ROC curves of anomaly detection performance using different algorithms on images from Air Force.
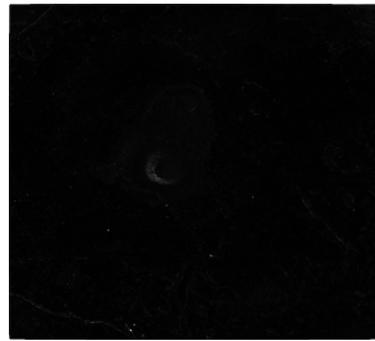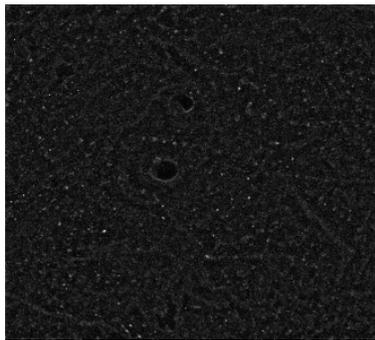

Figure 5.11: Highest detection rates and highest false alarm rates obtained from different algorithms on the second test data.
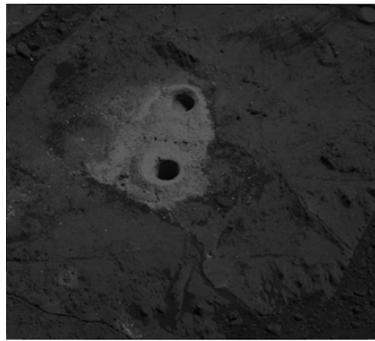
(a) Mastcam data
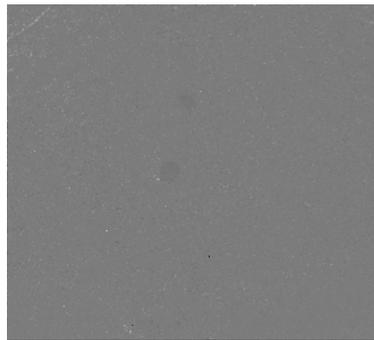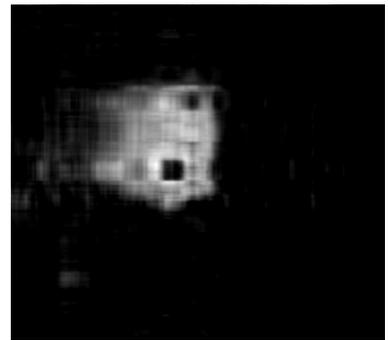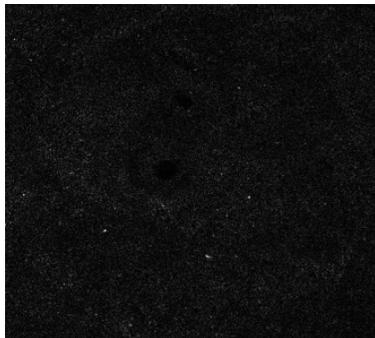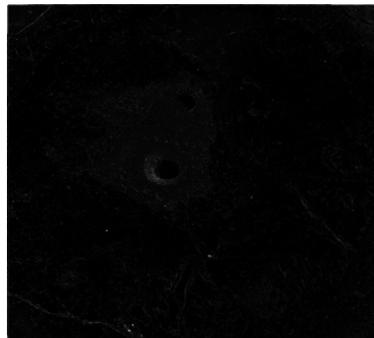
(b) GRX[123]

(c) LRX [142]

(d) SSRX [128]

(e) Li's [97]

(f) CRD [96]

(g) LSMAD [173]

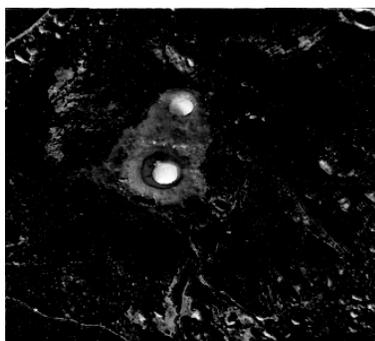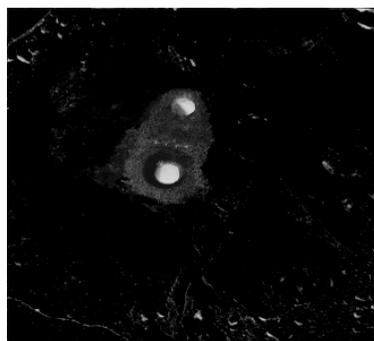**(h)** RPCA [160]

(i) Wang's [154]

(j) LRR [99]
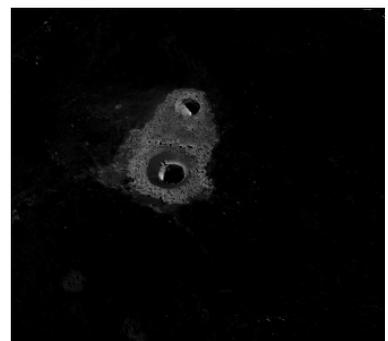
(k) ADLR

Figure 5.12: Detection results on the Mastcam multispectral image data.

(a)   (b)   (c)
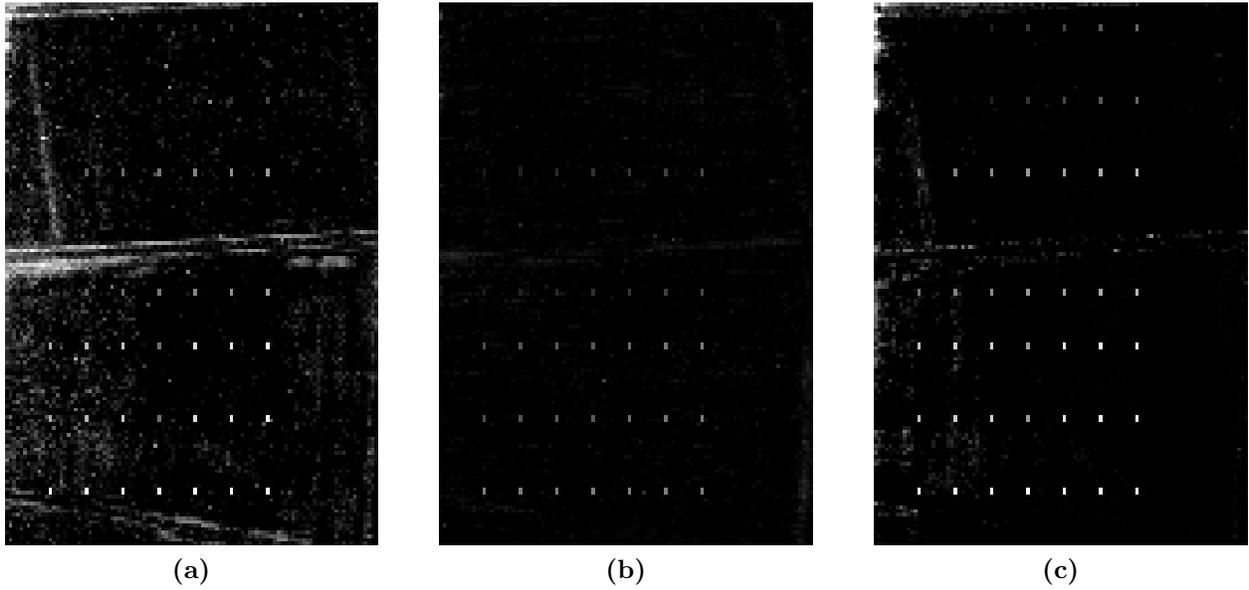
Figure 5.13: The detection results using different unmixing algorithms on the first test image. (a) Result generated by applying MLNMF [122]. (b) Result generated by applying R-CoNMF [95] (c) Result generated by applying MVC-NMF [109].

Figure 5.14: ROC curves of anomaly detection performance with different unmixing algorithms on the first test image.

Figure 5.15: Highest detection reates and highest false alarm rates obtained with different unmixing algorithms on the first test image.

(a)   **(b)**   **(c)**

Figure 5.16: The detection results using different unmixing algorithms on the second test image shown in Fig. 5.5a. (a) Result generated by applying MLNMF [122]. (b) Result generated by applying R-CoNMF [95]. (c) Result generated by applying MVC-NMF [109].
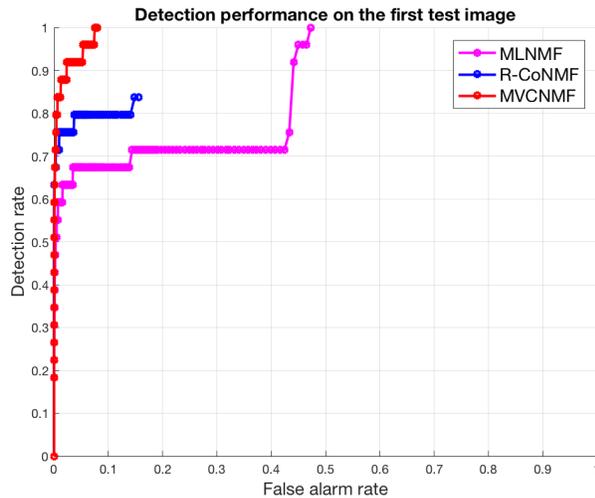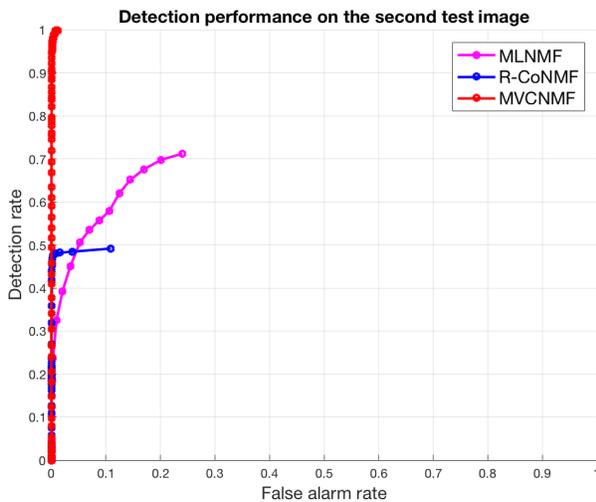

Figure 5.17: ROC curves of anomaly detection performance with different unmixing algorithms on the synthetic image.
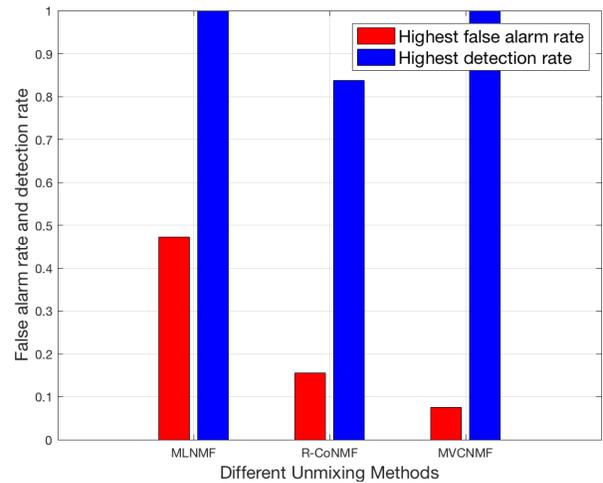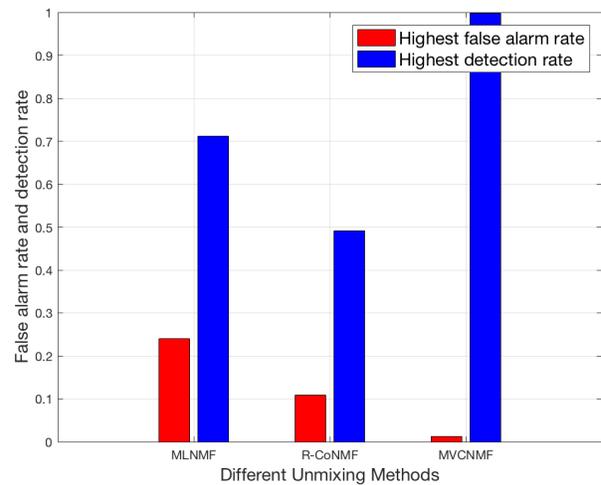

Figure 5.18: Highest detection reates and highest false alarm rates obtained with different unmixing algorithms on the second test image.
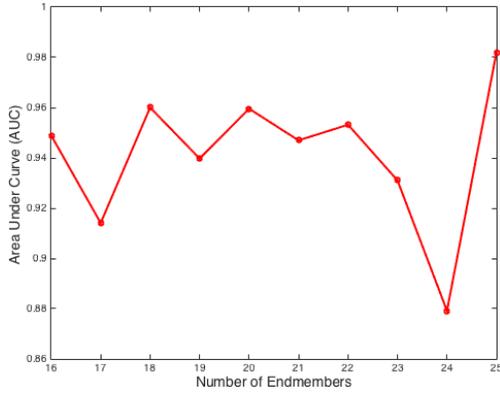
(a)
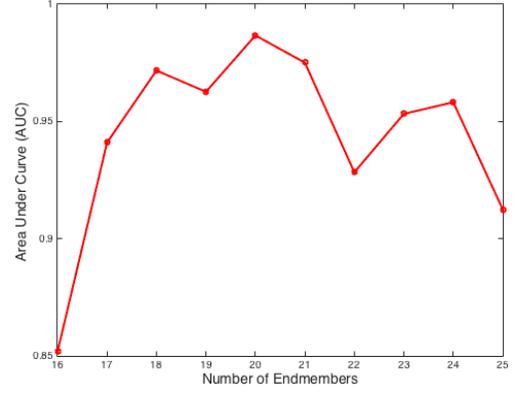
(b)

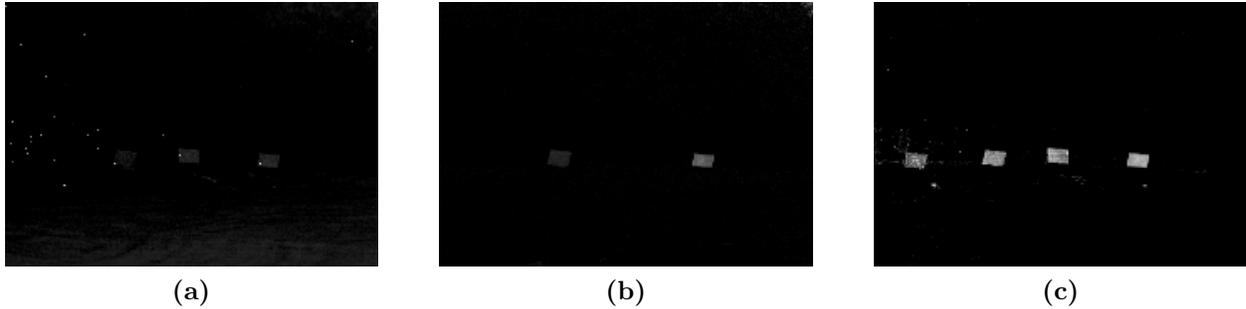**Figure 5.19:** The AUC of the proposed ADRL using different numbers of endmembers. (a) AUC of test image 1. (b) AUC of test image 2.
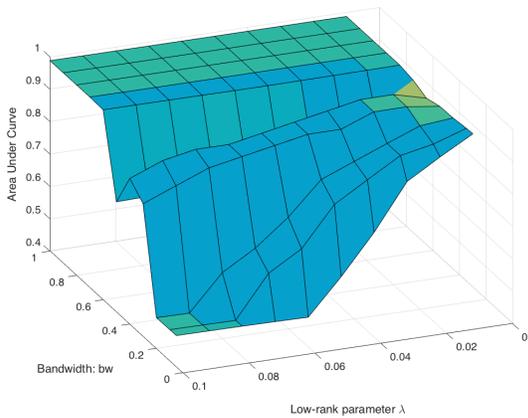
Figure 5.20: Parameter adjustment for test image 1.
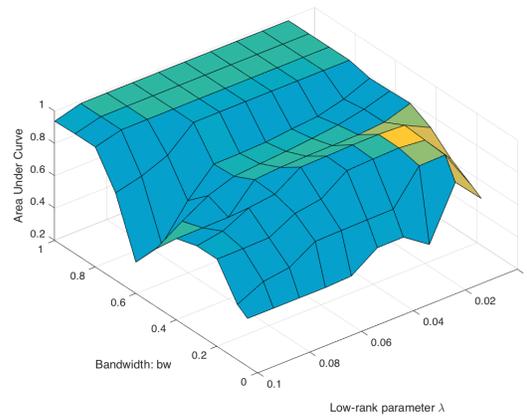
Figure 5.21: Parameter adjustment for test image 2.

Chapter 6

Summary and Future Work

This chapter summarizes the studies and discusses possible improvements for future research.

6.1   Summary

Hyperspectral images (HSI) have been widely utilized in computer vision and remote sensing fields and attracts a increasing attention in various domains. However, there are two challenge issues associated with HSI, *i.e.*, the existence of mixed pixels and its significantly low spatial resolution. To achieve better analysis results, in this dissertation, we proposed unsupervised spectral unmixing and hyperspectral image super-resolution approaches to address these two issues through deep learning, respectively. In addition, we improve the anomaly detection by projecting the data into subspaces through unmixing where anomalies behave differently from the background. The summary of the proposed work are elaborated below.

First, we proposed an untied denoising autoencoder with sparsity l_{21}, uDAS, to solve the unsupervised unmixing problem for highly noisy hyperspectral images. Different from existing part-based autoencoders, in uDAS, the encoder and decoder are untied and only the decoder is enforced to be non-negative. The proposed method, also powered with the denoising and l_{21} constraints, is able to increase the accuracy of the extracted endmembers, reduce redundant endmembers and the reconstruction error

simultaneously. With such properties, the proposed approach outperformed several state-of-the-art approaches, especially for highly noisy HSI data.

Second, we proposed an unsupervised sparse Dirichelet-Net (uSDN) to solve the problem of hyperspectral image super-resolution (HSI-SR). To the best of our knowledge, this is the first effort to solving the problem of HSI-SR in an unsupervised fashion. The network extracts the spectral basis from LR HSI with rich spectral information and spatial representations from HR MSI with high spatial information through a shared decoder. The representations from two modalities are encouraged to follow a sparse Dirichlet distribution. In addition, the angular difference of two representations is minimized during the optimization to reduce spectral distortion. Extensive experiments on two benchmark datasets demonstrate the superiority of the proposed approach over state-of-the-art.

Third, we proposed a novel anomaly detection algorithm based on performing low-rank decomposition on the abundance vectors from spectral unmixing and background coefficients from dictionary construction. In order to better separate the anomaly from background (especially isolated noisy pixels), we first performed spectral unmixing where the abundance matrix is estimated for further study instead of the raw data. To further remove noise and better represent the background as well as the anomaly, we constructed a dictionary based on the results from the mean-shift clustering algorithm. Therefore, the scattered noisy pixels can be clustered into the same group so as to lose the sparse property. Finally we proposed a low-rank decomposition method to detect the anomaly. Instead of directly encouraging the background to be low-rank, we applied the low-rank constraint on the coefficient of the background which is highly mixed. Experiments on three testing datasets showed that our algorithm is able to achieve high detection rate while maintaining low false alarm rate regardless of the type of the input images.

6.2 Future Works

We would like to improve our work in two aspects.

First, although our work filled the gap between deep learning and unsupervised spectral unmixing, like previous spectral unmixing works, it also requires carefully initialization

to avoid being tracked by local minimum. Deep learning is generally solved by back-propagation, which was first applied to solve neural networks in the 1970s. However, back-propagation is not guaranteed to find the global minimum of the error function. We amended this by carefully initialize the network weights using vortex component analysis. Although deep learning usually encourages its nodes to be independent from one another, we found that the interdependency of the nodes may stabilizes the network to a large extend such that no specific initialization procedure is needed. In the future work, we would like try to decrease the probability of achieving a local minimum during the optimization by take advantage of the interdependency between nodes.

Second, we made the first effort to solve the hyperspectral super-resolution (HSI-SR) problem with unsupervised deep learning. In this work, a brand new network architecture was designed to fuse two different modalities, *i.e.*, low spatial resolution (LR) HSI and high spatial resolution (HR) MSI. And our goal is to generate images of high resolution in both the spatial or spectral domains, *i.e.*, HR HSI. It is worth to mention that the target images, HR HSI and two source images LR HSI and HR MSI, are three modalities laying in different domains with different dimensions. The key to solve this problem is to take advantage of the shared information among them. It is very close to the problem of transfer learning, where the knowledge gained from a source domain can be adopted for a target domain, *e.g.*, the network trained on printed digits can be adopted to recognize handwritten digits. The similarity between the HSI-SR problem and the transfer learning problem is that there exists shared information between different domains. Therefore, in the future work, we would like to generalize our proposed network by adopting it on transfer learning problems.

Bibliography

[1] Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322. 69

[2] Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., and Selva, M. (2006). Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596. 15, 46, 57

[3] Aiazzi, B., Baronti, S., and Selva, M. (2007). Improving component substitution pansharpening through multivariate regression of ms+ pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10). 15, 46, 57

[4] Akhtar, N., Shafait, F., and Mian, A. (2015). Bayesian sparse representation for hyperspectral image super resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3631–3640. 3, 4, 15, 16, 46, 47, 57, 58, 59

[5] Akhtar, N., Shafait, F., and Mian, A. (2016). Hierarchical beta process with gaussian process prior for hyperspectral image super resolution. *European Conference on Computer Vision*, pages 103–120. 57, 58

[6] Altmann, Y., McLaughlin, S., and Hero, A. (2015). Robust linear spectral unmixing using anomaly detection. *IEEE Transactions on Computational Imaging*, 1(2):74–85. 2, 20

[7] Andreou, C. and Karathanassi, V. (2014). Estimation of the number of endmembers using robust outlier detection method. *IEEE journal of selected topics in Applied earth observations and Remote sensing*, 7(1):247–256. 30

[8] Ayhan, B., Dao, M., Kwan, C., Chen, H. M., Bell, J. F., and Kidd, R. (2017). A novel utilization of image registration techniques to process mastcam images in mars rover with applications to image fusion, pixel clustering, and anomaly detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP(99):1–12. 77

[9] Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. *ICML unsupervised and transfer learning*, 27:1. 32

[10] Bengio, Y. et al. (2009). Learning deep architectures for ai. *Foundations and trends®️ in Machine Learning*, 2(1):1–127. 21

[11] Bertsekas, D. P. (2014). *Constrained optimization and Lagrange multiplier methods.* Academic press. 33

[12] Bioucas-Dias, J. and Nascimento, J. (2008a). Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8). 30, 40

[13] Bioucas-Dias, J. and Nascimento, J. (2008b). Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445. 84

[14] Bioucas-Dias, J. M. (2009). A variable splitting augmented lagrangian approach to linear spectral unmixing. pages 1–4. IEEE. 12

[15] Bioucas-Dias, J. M. and Figueiredo, M. A. (2010). Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, pages 1–4. 67

[16] Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N. M., and Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. *Geoscience and Remote Sensing Magazine, IEEE*, 1(2):6–36. 3, 65

[17] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(2):354–379. 1, 2, 4, 11, 13, 20, 22, 27, 65, 67

[18] Borengasser, M., Hungate, W. S., and Watkins, R. (2007). *Hyperspectral remote sensing: principles and applications.* Crc Press. 1, 2, 65

[19] Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540. 15

[20] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11. 19

[21] Chaib, S., Liu, H., Gu, Y., and Yao, H. (2017). Deep feature fusion for vhr remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing.* 3

[22] Chakrabarti, A. and Zickler, T. (2011). Statistics of real-world hyperspectral images. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 193–200. 1, 57

[23] Chang, C.-I. and Chiang, S.-S. (2002). Anomaly detection and classification for hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 40(6):1314–1325. 65

[24] Chang, C.-I. and Liu, K.-H. (2014). Progressive band selection of spectral unmixing for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(4):2002–2017. 67

[25] Chavez, P., Sides, S. C., and Anderson, J. A. (1991). Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 57(3). 15, 46

[26] Chen, M., Weinberger, K., Sha, F., and Bengio, Y. (2014). Marginalized denoising auto-encoders for nonlinear representations. *Proceedings of The 31st International Conference on Machine Learning*, pages 1476–1484. 10, 29

[27] Chen, S., Wang, H., Xu, F., and Jin, Y.-Q. (2016). Target classification using the deep convolutional networks for sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4806–4817. 3

[28] Chen, Y., M, N. N., and Tran, T. D. (2011a). Sparse representation for target detection in hyperspectral imagery. *Selected Topics in Signal Processing, IEEE Journal of*, 5(3):629–640. 7, 18

[29] Chen, Y., Nasrabadi, N. M., and Tran, T. D. (2011b). Simultaneous joint sparsity model for target detection in hyperspectral imagery. *Geoscience and Remote Sensing Letters, IEEE*, 8(4):676–680. 18

[30] Chen, Y., Nasrabadi, N. M., and Tran, T. D. (2013). Hyperspectral image classification via kernel sparse representation. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):217–231. 18

[31] Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., and Pan, C. (2017). Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3322–3337. 3

[32] Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415. 3

[33] Cheriyadat, A. and Bruce, L. M. (2003). Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, 6:3420–3422 vol.6. 7

[34] Chorowski, J. and Zurada, J. M. (2015). Learning understandable neural networks with nonnegative weight constraints. *IEEE transactions on neural networks and learning systems*, 26(1):62–69. 14

[35] Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619. 69

[36] Dao, M., Kwan, C., Ayhan, B., and Bell, J. F. (2017). Enhancing mastcam images for mars rover mission. *Advances in Neural Networks - ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II*, pages 197–206. 77

[37] Dian, R., Fang, L., and Li, S. (2017). Hyperspectral image super-resolution via non-local sparse tensor factorization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5353. 3, 15, 46, 47, 49

[38] Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J.-Y., and Hero, A. O. (2009a). Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery. *IEEE Transactions on Signal Processing*, 57(11). 13, 35, 39

[39] Dobigeon, N., Moussaoui, S., Tourneret, J.-Y., and Carteret, C. (2009b). Bayesian separation of spectral sources under non-negativity and full additivity constraints. *Signal Processing*, 89(12):2657–2669. 13

[40] Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*. 10

[41] Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Image super-resolution using deep convolutional networks. *arXiv preprint arXiv:1501.00092*. 17

[42] Dong, C., Loy, C. C., He, K., and Tang, X. (2016a). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307. 5, 47

[43] Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., and Li, X. (2016b). Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, 25(5):2337–2352. 16, 50

[44] Du, B. and Zhang, L. (2011). Random-selection-based anomaly detector for hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(5):1578–1589. 65

[45] Du, B. and Zhang, L. (2014). A discriminative metric learning based anomaly detection method. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(11):6844–6857. 77

[46] Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. (2001). Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, pages 472–478. 53

[47] Eches, O., Dobigeon, N., and Tourneret, J.-Y. (2010). Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical bayesian algorithm. *IEEE Journal of Selected Topics in Signal Processing*, 4(3):582–591. 30

[48] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 681–687. 11

[49] Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J., and Tilton, J. C. (2013). Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675. 1

[50] Fu, Y., Zheng, Y., Sato, I., and Sato, Y. (2016). Exploiting spectral-spatial correlation for coded hyperspectral image restoration. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1

[51] Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. on Information Theory*, IT-21(1):32–40. 70

[52] G. Swayze, S. S. and Gallagher, A. (1992). Aviris mineral mapping at cuprite, nevada. *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1:47–49. 34, 35

[53] Gong, M., Zhan, T., Zhang, P., and Miao, Q. (2017). Superpixel-based difference representation learning for change detection in multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2658–2673. 3

[54] Goodfellow, I., Bengio, Y., and Courville, A. (2016a). *Deep learning*. MIT press. 9

[55] Goodfellow, I., Bengio, Y., and Courville, A. (2016b). *Deep Learning*. MIT Press. http://www.deeplearningbook.org. 21, 53

[56] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout networks. *arXiv preprint arXiv:1302.4389*. 26

[57] Gülçehre, Ç. and Bengio, Y. (2016). Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research*, 17(8). 33

[58] Guo, R., Wang, W., and Qi, H. (2015). Hyperspectral image unmixing using autoencoder cascade. *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2015 7th Workshop on*, pages 1–4. 14, 67

[59] Han, J. and Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. *From Natural to Artificial Neural Computation*, pages 195–201. 53

[60] Han, J., Zhang, D., Cheng, G., Guo, L., and Ren, J. (2015). Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337. 3

[61] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108. 69

[62] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 5, 17, 47, 48

[63] Heinz, D. C. et al. (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 39(3):529–545.

[64] Heylen, R., Akhter, M. A., and Scheunders, P. (2013). On using projection onto convex sets for solving the hyperspectral unmixing problem. *IEEE Geoscience and Remote Sensing Letters*, 10(6):1522–1526. 40

[65] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554. 9

[66] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507. 11

[67] Hosseini-Asl, E., Zurada, J. M., and Nasraoui, O. (2016). Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints. *IEEE transactions on neural networks and learning systems*, 27(12):2486–2498. 14

[68] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2016). Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*. 53

[69] Huang, S. and Tran, T. D. (2017). Sparse signal recovery via generalized entropy functions minimization. *arXiv preprint arXiv:1703.10556*. 53

[70] Huang, W., Xiao, L., Wei, Z., Liu, H., and Tang, S. (2015). A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1037–1041. 3, 17, 47

[71] Iordache, M.-D., Bioucas-Dias, J. M., and Plaza, A. (2011). Sparse unmixing of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6):2014–2039. 21, 30

[72] Iordache, M.-D., Bioucas-Dias, J. M., and Plaza, A. (2012). Total variation spatial regularization for sparse hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4484–4502. 40

[73] Iordache, M.-D., Bioucas-Dias, J. M., and Plaza, A. (2014). Collaborative sparse regression for hyperspectral unmixing. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(1):341–354. 67

[74] Jiang, Z., Lin, Z., and Davis, L. S. (2011). Learning a discriminative dictionary for sparse coding via label consistent k-svd. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1697–1704. 69

[75] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, pages 694–711. 17

[76] Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., and Schindler, K. (2017). Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing.* 3

[77] Kawakami, R., Matsushita, Y., Wright, J., Ben-Ezra, M., Tai, Y.-W., and Ikeuchi, K. (2011). High-resolution hyperspectral imaging via matrix factorization. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2329–2336. 16, 46, 49, 50, 58

[78] Kemker, R. and Kanan, C. (2017). Self-taught feature learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2693–2705. 3

[79] Keshava, N. (2003). A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal*, 14(1):55–78. 2, 4, 20, 27

[80] Khan, S. H., He, X., Porikli, F., and Bennamoun, M. (2017). Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing.* 3

[81] Kim, J., Kwon Lee, J., and Mu Lee, K. (2016a). Accurate image super-resolution using very deep convolutional networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5, 47

[82] Kim, J., Kwon Lee, J., and Mu Lee, K. (2016b). Deeply-recursive convolutional network for image super-resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5, 47

[83] Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2):79–88. 52

[84] Kwan, C., Ayhan, B., Chen, G., Wang, J., Ji, B., and Chang, C.-I. (2006). A novel approach for spectral unmixing, classification, and concentration estimation of chemical and biological agents. *IEEE Transactions on Geoscience and Remote Sensing*, 44(2):409–419. 1

[85] Kwon, H. and Nasrabadi, N. M. (2005). Kernel matched signal detectors for hyperspectral target detection. *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 6–6. 2

[86] Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 47

[87] Lanaras, C., Baltsavias, E., and Schindler, K. (2015). Hyperspectral super-resolution by coupled spectral unmixing. *Proceedings of the IEEE International Conference on Computer Vision*, pages 3586–3594. 3, 4, 16, 46, 47, 49, 50, 51, 57, 58, 59

[88] Landgrebe, D. (2002). Hyperspectral image data analysis. volume 19, pages 17–28. IEEE. 65

[89] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. 9

[90] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551. 3

[91] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*. 5, 17, 47

[92] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. 7

[93] Lemme, A., Reinhart, F., and Steil, J. J. (2010). Efficient online learning of a non-negative sparse autoencoder. *European Symposium Artificial Neural Networks*. 14, 22

[94] Li, J., Agathos, A., Zaharie, D., Bioucas-Dias, J. M., Plaza, A., and Li, X. (2015). Minimum volume simplex analysis: A fast algorithm for linear hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 53(9). 12, 35, 39, 40

[95] Li, J., Bioucas-Dias, J. M., Plaza, A., and Liu, L. (2016). Robust collaborative nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6076–6090. xiv, 13, 27, 33, 34, 35, 40, 67, 68, 83, 90, 91

[96] Li, W. and Du, Q. (2015). Collaborative representation for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1463–1474. 19, 78, 86, 87, 89

[97] Li, Shuangjiang, W. W. Q. H. A. B. K. C. and Vance, S. (2015). Low-rank tensor decomposition based anomaly detection for hyperspectral imagery. pages 4525–4529. IEEE. 19, 66, 78, 86, 87, 89

[98] Lin, Z., Chen, M., and Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*. 19, 66, 72, 74

[99] Liu, G., Lin, Z., and Yu, Y. (2010). Robust segmentation by low-rank representation. *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670. 19, 66, 72, 73, 74, 75, 86, 87, 89

[100] Loncan, L., de Almeida, L. B., Bioucas-Dias, J. M., Briottet, X., Chanussot, J., Dobigeon, N., Fabre, S., Liao, W., Licciardi, G. A., Simoes, M., et al. (2015). Hyperspectral pansharpening: a review. *IEEE Geoscience and remote sensing magazine*, 3(3). 4, 15, 47, 49, 54, 57

[101] Lu, G. and Fei, B. (2014). Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901–010901. xi, 1

[102] Lu, J. and Forsyth, D. (2015). Sparse depth super resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 47

[103] Maggiori, E., Charpiat, G., Tarabalka, Y., and Alliez, P. (2017). Recurrent neural networks to correct satellite image classification maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4962–4971. 1, 3

[104] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7). 15

[105] Manolakis, D. and Shaw, G. (2002). Detection algorithms for hyperspectral imaging applications. *Signal Processing Magazine, IEEE*, 19(1):29–43. 65

[106] Marrero, R., Lopez, S., Callicó, G. M., Veganzones, M. A., Plaza, A., Chanussot, J., and Sarmiento, R. (2015). A novel negative abundance-oriented hyperspectral unmixing algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3772–3790. 67

[107] Masi, G., Cozzolino, D., Verdoliva, L., and Scarpa, G. (2016). Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7). 17, 47

[108] Matteoli, S., Diani, M., and Corsini, G. (2010). A tutorial overview of anomaly detection in hyperspectral images. *IEEE Aerospace and Electronic Systems Magazine*, 25(7):5–28. 3

[109] Miao, L. and Qi, H. (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(3):765–777. xi, xiv, 12, 13, 19, 34, 35, 36, 39, 40, 67, 69, 90, 91

[110] Miao, L., Qi, H., and Szu, H. (2007). A maximum entropy approach to unsupervised mixed-pixel decomposition. *Image Processing, IEEE Transactions on*, 16(4):1008–1021. 13, 67

[111] Nalisnick, E. and Smyth, P. (2017). Deep generative models with stick-breaking priors. *ICML*. 52

[112] Nascimento, J. M. and Bioucas-Dias, J. M. (2012). Hyperspectral unmixing based on mixtures of dirichlet components. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3). 14

[113] Nascimento, J. M. and Dias, J. M. (2005). Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910. 12, 33

[114] Nasrabadi, N. M. (2008). Regularization for spectral matched filter and rx anomaly detector. *SPIE Defense and Security Symposium*, pages 696604–696604. 18

[115] Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, pages 1–19. 10

[116] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696. 11

[117] Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). Efficient and robust feature selection via joint 2, 1-norms minimization. *Advances in neural information processing systems*, pages 1813–1821. 31, 33

[118] Ozkan, S., Kaya, B., Esen, E., and Akar, G. (2017a). *EndNet: Sparse AutoEncoder Network for Endmember Extraction and Hyperspectral Unmixing.* 54

[119] Ozkan, S., Kaya, B., Esen, E., and Akar, G. B. (2017b). Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing. *arXiv preprint arXiv:1708.01894.* 15

[120] Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., et al. (2009). Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:S110–S122. 3

[121] Plaza, A., Du, Q., Bioucas-Dias, J. M., Jia, X., and Kruse, F. A. (2011). Foreword to the special issue on spectral unmixing of remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4103–4110. 2

[122] Rajabi, R. and Ghassemian, H. (2015). Spectral unmixing of hyperspectral imagery using multilayer nmf. *IEEE Geoscience and Remote Sensing Letters*, 12(1):38–42. xiv, 13, 35, 40, 67, 68, 83, 90, 91

[123] Reed, I. S. and Yu, X. (1990). Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. volume 38, pages 1760–1770. IEEE. 18, 78, 86, 87, 89

[124] Rogge, D. M., Rivard, B., Zhang, J., and Feng, J. (2006). Iterative spectral unmixing for optimizing per-pixel endmember sets. *IEEE Transactions on Geoscience and Remote Sensing*, 44(12):3725–3736. 21

[125] Romero, A., Gatta, C., and Camps-Valls, G. (2016). Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362. 3

[126] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, DTIC Document. 10

[127] Sarker, C. D., Jia, X., Wang, L., Fraser, D., and Lymburner, L. (2015). Spectral unmixing with estimated adaptive endmember index using extended support vector machine. *Spatial Diversity and Dynamics in Resources and Urban Development*, pages 37–71. 67

[128] Schaum, A. (2004). Joint subspace detection of hyperspectral targets. volume 3, page 1824 Vol.3. 18, 78, 86, 87, 89

[129] Schweizer, S. M. and Moura, J. M. (2001). Efficient detection in hyperspectral imagery. volume 10, pages 584–597. IEEE. 77

[130] Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650. 52

[131] Shensa, M. J. (1992). The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on signal processing*, 40(10):2464–2482. 15

[132] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 47

[133] Shi, Z. and Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3623–3634. 3

[134] Sigurdsson, J., Ulfarsson, M. O., and Sveinsson, J. R. (2014). Hyperspectral unmixing with l_{q} regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):6793–6806. 40

[135] Simões, M., Bioucas-Dias, J., Almeida, L. B., and Chanussot, J. (2015). A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6). 16, 49, 57

[136] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.* 17

[137] Snyder, W. E. and Qi, H. (2017). *Fundamentals of Computer Vision.* Cambridge University Press. 70

[138] Somers, B., Asner, G. P., Tits, L., and Coppin, P. (2011). Endmember variability in spectral mixture analysis: A review. *Remote Sensing of Environment*, 115(7):1603–1616. 30

[139] Spangler, L. H., Dobeck, L. M., Repasky, K. S., Nehrir, A. R., Humphries, S. D., Barr, J. L., Keith, C. J., Shaw, J. A., Rouse, J. H., Cunningham, A. B., et al. (2010). A shallow subsurface controlled release facility in bozeman, montana, usa, for testing near surface co2 detection techniques and transport models. *Environmental Earth Sciences*, 60(2):227–239. 1

[140] Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, pages 2222–2230. 11

[141] Stein, D. W., Beaven, S. G., Hoff, L. E., Winter, E. M., Schaum, A. P., and Stocker, A. D. (2002). Anomaly detection from hyperspectral imagery. *Signal Processing Magazine, IEEE*, 19(1):58–69. 65

[142] Taitano, Y. P., Geier, B. A., and Bauer, K. W. (2010). A locally adaptable iterative rx detector. volume 2010, page 11. Hindawi Publishing Corp. 18, 78, 86, 87, 89

[143] Tang, J., Deng, C., Huang, G.-B., and Zhao, B. (2015). Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1174–1185. 3

[144] Theiler, J. and Wohlberg, B. (2012). Local coregistration adjustment for anomalous change detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(8):3107–3116. 65

[145] Thomas, C., Ranchin, T., Wald, L., and Chanussot, J. (2008). Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1301–1312. 15, 46, 47

[146] Uzkent, B., Hoffman, M. J., and Vodacek, A. (2016). Real-time vehicle tracking in aerial video using hyperspectral features. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1

[147] Uzkent, B., Rangnekar, A., and Hoffman, M. (2017). Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1

[148] Van der Meer, F. D. and Jia, X. (2012). Collinearity and orthogonality of endmembers in linear spectral unmixing. *International Journal of Applied Earth Observation and Geoinformation*, 18:491–503. 67

[149] Van Nguyen, H., Banerjee, A., and Chellappa, R. (2010). Tracking via object reflectance using a hyperspectral video camera. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 44–51. 1

[150] Veganzones, M. A., Simoes, M., Licciardi, G., Yokoya, N., Bioucas-Dias, J. M., and Chanussot, J. (2016). Hyperspectral super-resolution of locally low rank images from complementary multisource data. *IEEE Transactions on Image Processing*, 25(1):274–288. 16, 50

[151] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408. 10

[152] Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., Restaino, R., and Wald, L. (2015). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5). 2, 46, 49

[153] Volpi, M. and Tuia, D. (2017). Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893. 3

[154] W. Wang, S. Li, H. Q. B. A. C. K. and Vance, S. (2015). Identify anomaly component by sparsity and low rank. IEEE. 19, 66, 78, 79, 86, 87, 89

[155] Wang, H., Raj, B., and Xing, E. P. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*. 3, 9

[156] Wei, Q., Bioucas-Dias, J., Dobigeon, N., and Tourneret, J.-Y. (2015). Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7). 16, 49, 57, 58

[157] Wei, Q., Dobigeon, N., and Tourneret, J.-Y. (2014). Bayesian fusion of hyperspectral and multispectral images. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 16

[158] Wei, Y., Yuan, Q., Shen, H., and Zhang, L. (2017). Boosting the accuracy of multi-spectral image pan-sharpening by learning a deep residual network. *arXiv preprint arXiv:1705.07556*. 17, 48

[159] Winter, M. E. (1999). N-findr: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. 12

[160] Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, pages 2080–2088. 74, 86, 87, 89

[161] Wycoff, E., Chan, T.-H., Jia, K., Ma, W.-K., and Ma, Y. (2013). A non-negative sparse promoting algorithm for high resolution hyperspectral imaging. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 1409–1413. 16, 51

[162] Xu, Y., Wu, Z., Li, J., Plaza, A., and Wei, Z. (2016). Anomaly detection in hyperspectral images based on low-rank and sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4):1990–2000. 69

[163] Yao, X., Han, J., Cheng, G., Qian, X., and Guo, L. (2016). Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3660–3671. 3

[164] Yasuma, F., Mitsunaga, T., Iso, D., and Nayar, S. K. (2010). Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253. 57

[165] Yokoya, N., Grohnfeldt, C., and Chanussot, J. (2017). Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56. 4, 47, 54

[166] Yokoya, N., Yairi, T., and Iwasaki, A. (2012). Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537. 16, 49, 50, 51, 57, 58

[167] Yu, Y., Li, J., Guan, H., and Wang, C. (2016). Automated detection of three-dimensional cars in mobile laser scanning point clouds using dbm-hough-forests. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4130–4142. 3

[168] Zhang, B., Zhuang, L., Gao, L., Luo, W., Ran, Q., and Du, Q. (2014a). Pso-em: a hyperspectral unmixing algorithm based on normal compositional model. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12):7782–7792. 67

[169] Zhang, F., Du, B., and Zhang, L. (2016a). Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1793–1802. 1, 3

[170] Zhang, L., Du, B., and Zhong, Y. (2010). Hybrid detectors based on selective endmembers. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(6):2633–2646. 65, 67

[171] Zhang, L., Zhang, L., Tao, D., Huang, X., and Du, B. (2014b). Hyperspectral remote sensing image subpixel target detection based on supervised metric learning. *IEEE transactions on geoscience and remote sensing*, 52(8):4955–4965. xi, 2

[172] Zhang, X., Zhang, A., and Meng, X. (2015). Automatic fusion of hyperspectral images and laser scans using feature points. *Journal of Sensors*, 2015. 3

[173] Zhang, Y., Du, B., Zhang, L., and Wang, S. (2016b). A low-rank and sparse matrix decomposition-based mahalanobis distance method for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1376–1389. 19, 66, 78, 86, 87, 89

[174] Zhao, R., Du, B., and Zhang, L. (2017a). Hyperspectral anomaly detection via a sparsity score estimation framework. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3208–3222. 3

[175] Zhao, W. and Du, S. (2016). Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4544–4554. 3

[176] Zhao, W., Wang, Z., Gong, M., and Liu, J. (2017b). Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network. *IEEE Transactions on Geoscience and Remote Sensing*. 3

[177] Zhou, J., Kwan, C., and Ayhan, B. (2017). Improved target detection for hyperspectral images using hybrid in-scene calibration. *J. Appl. Remote Sens*, 11(3). 77

[178] Zhou, J., Kwan, C., Ayhan, B., and Eismann, M. T. (2016a). A novel cluster kernel rx algorithm for anomaly and change detection using hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11):6497–6504. 77

[179] Zhou, J., Kwan, C., and Budavari, B. (2016b). Hyperspectral image super-resolution: a hybrid color mapping approach. *Journal of Applied Remote Sensing*, 10(3):035024–035024. 4, 47

[180] Zhou, T. and Tao, D. (2011). Godec: Randomized low-rank & sparse matrix decomposition in noisy case. 19, 72, 74, 78

Vita

Ying Qu was born in Datong, Shanxi province, P.R. China. She received her Bachelor and Master degrees in 'Automation' and 'Pattern Recognition and Intelligent Systems' from Northeastern University (China) in 2008 and 2010, respectively. From fall 2012, she enrolled into the doctoral program at the University of Tennessee at Knoxville in the department of Electrical Engineering and Computer Science. She joined the Advanced Imaging and Collaborative Information Processing (AICIP) lab as a graduate research assistant under supervision of Professor Hairong Qi in November 2015. Her major research areas are image processing, remote sensing, computer vision and pattern recognition.