Normalizing Flow-Based Probability Distribution Representation Detector for Hyperspectral Anomaly Detection

Xiaorun Li, Shaoqi Yu, Shuhan Chen, *Member, IEEE*, and Liaoying Zhao

***Abstract*—Due to the powerful reconstruction ability, deep learning based hyperspectral anomaly detection methods have been prevalent in recent years. However, the capability of neural networks and the meaning of latent space remains unexplainable to some extent. To address the issue, we propose a normalizing flow-based probability distribution representation detector (NF-PDRD) for hyperspectral anomaly detection in this article, which clarifies the capability of the model from a probabilistic perspective. The framework first utilizes the variational autoencoder to acquire the probability distribution representation with the mean vector and standard deviation vector for the original data. Subsequently, we introduce a normalizing flow to transform the Gaussian approximate posterior to a more complex distribution, making the model generative and expressive. We finally accomplish the detection process with the extracted probabilistic representation data using the strategy of Gaussian mixture model estimation to fully leverage the spatial information. Experimental results on both synthetic and real data sets demonstrate the outstanding performance of the proposed NF-PDRD.**

***Index Terms*—Anomaly detection, hyperspectral imagery (HSI), normalizing flows, probability representation, variational autoencoder (VAE).**

I. Introduction

DUE to the rich spectral knowledge about the characteristics of materials, hyperspectral imagery (HSI) plays an essential role in remote sensing applications [1]. The spectrum of each pixel represents the radiance or reflectance value at the corresponding band covering the range of wavelengths from visible to near-infrared or short-wave infrared [2]. Relying on these properties, HSIs have been applied in various fields, such as material exploration [3], precise agriculture [4], civilian rescue [5], and environmental monitoring [6]. Hyperspectral anomaly detection can be seen as a binary classification problem that aims to locate the pixels whose spectral signatures are significantly different from their local surroundings [7], which can be regarded as a special case of hyperspectral classification [8], [9]. Different from target detection, anomaly detection is a very challenging and promising task without any prior knowledge [10]. The anomalies we studied here are defined objects in small regions whose spectral signature significantly differs from the neighboring areas.

Over the past several decades, researchers have proposed a lot of effective hyperspectral anomaly detection algorithms. One of the most well-known anomaly detection methods is the Reed-Xiaoli (RX) detector [11]. It develops from the generalized likelihood ratio test and assumes that the background obeys a multivariate Gaussian distribution. The RX detector has two versions: global RX uses the global information, and local RX uses local information to estimate the background statistics. However, the existing noise and contamination of adjacent pixels make it hard to satisfy the assumption in real scenarios [12]. Therefore, a series of algorithms improve the performance based on the RX detector. The weighted RX can handle the covariance corruption problem by reducing the weight of the anomalies or noise and increasing the weight of the background samples [13]. By dividing the pixels into several clusters, the cluster-based anomaly detection method detects the anomalies based on each cluster [14]. The dual window-based eigen separation transform detector can enhance the discrimination between the anomalies and the background in a low-dimensional subspace [15]. In addition, a lot of non-Gaussian models have been proposed. The kernel RX (KRX) algorithm [16] projects the original image into a nonlinear high-dimensional feature space, which considers high order characteristics of different feature bands. The cluster KRX [17] applies a clustering process before detection, which significantly reduces the computational cost.

In recent years, various collaborative representation and matrix decomposition based methods have emerged due to the development of compressive sensing theory [18]. Li *et al.* [19] present a collaborative representation detector (CRD) based on the concept that each background pixel can be seen as a linear combination of its adjacent pixels, while anomalies cannot. A low-rank and sparse matrix decomposition [20] exploits the robust principal component analysis [21] strategy and decomposes the image into the background component and anomaly component. The underlying hypothesis is that all pixels lie in the same subspace, which is usually unreasonable

Manuscript received February 15, 2022; revised May 18, 2022; accepted June 8, 2022. Date of publication June 13, 2022; date of current version June 24, 2022. This work was supported in part by the National Nature Science Foundation of China under Grant 62171404, and in part by the Joint Fund of the Ministry of Education of China under Grant 8091B022118. *(Corresponding author: Shaoqi Yu.)*

Xiaorun Li, Shaoqi Yu, and Shuhan Chen are with the Department of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: lxrly@zju.edu.cn; rocket.ysq@gmail.com; 11410057@zju.edu.cn).

Liaoying Zhao is with the Department of Computer Science, Hangzhou Dianzi University, Hangzhou 310027, China (e-mail: zhaoly@hdu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3182538

in real scenes. To solve the problem, the low-rank and sparse representation (LRASR) [22] algorithm introduces the concept of background dictionary to project the image into multiple subspaces. Yu *et al.* [23] use the local outlier factor to construct the background dictionary and design a matched filter to enhance the discrimination between the anomalies and the background. A low-rank and sparse matrix decomposition based Mahalanobis distance (LSMAD) [24] method takes the effect of noise into consideration when extracting the anomalies from the background component. A low-rank and sparse decomposition model with a mixture of Gaussian (LSDM-MoG) [25] characterizes the data with multiple complex distributions by using variational Bayes (VB). Moreover, the total variation and sparsity regularized decomposition model [26] incorporates the total variation term with sparsity-inducing regularizations to facilitate the separation and presents a novel technique to construct the dictionary with density peak based clustering. These matrix decomposition based detectors can achieve superior performance compared to traditional algorithms. However, most of them only consider the spectral properties of anomaly components but ignore their spatial distribution features.

Nowadays, with the improvement of computational capability, a large number of deep learning based hyperspectral anomaly detection methods have been proposed, including convolutional neural network (CNN) [27], [28], deep belief network (DBN) [29], [30], autoencoder (AE) [31], [32], variational autoencoder (VAE) [33], adversarial autoencoder [34], [35], and generative adversarial network (GAN) [36], [37]. Li *et al.* [28] train a multilayer CNN using pixel pairs generated from the reference image. By introducing a suppression function, Lei *et al.* [29] propose a spectral–spatial feature extraction detector that can construct a discriminative feature space based on the DBN architecture. In literature [32], nonlinear feature extraction based on convolutional stacked autoencoder neural networks are implemented, and the statistical properties of the segmented areas are used to suppress the false alarms. Lei *et al.* [33] propose a novel discriminative reconstruction method with spectral learning that uses the spectral error map to detect anomalies and introduces the spectral angle distance to formulate the loss function. Xie *et al.* [34] propose a spectral adversarial feature learning anomaly detector that introduces spectral constraint loss and adversarial loss in the network with batch normalization to extract the intrinsic features. Jiang *et al.* [37] use GAN [38] to evaluate the background statistics and acquire a spectral feature in a novel semisupervised spectral learning. These deep learning based methods mainly focus on enhancing the discrimination between the anomalies and the background by using reconstruction error but hardly concern the probability distribution characteristics in the latent space, which may further influence the detection power and background suppression to some extent.

Due to the excellent reconstruction ability, VAE [39] has received considerable attention in the field of hyperspectral anomaly detection. Moreover, vanilla VAE owns the representation ability to extract the potential probability distribution for each training sample. The architecture assumes that the approximate posterior distribution of the latent variable conforms to a Gaussian distribution. However, it can be any complex distribution in real applications, which contradicts the potential assumption. To address the problem, we propose a normalizing flow-based probability distribution representation detector (NF-PDRD), which introduces normalizing flow when the VAE network generates the latent variable for each sample. The encoder first extracts an initial Gaussian distribution that serves as the basic distribution. Then, the normalizing flow module transforms the simple distribution into an unknown complex distribution by applying a sequence of invertible transformations until an expected level of complexity is satisfied. The decoder outputs the reconstructed data, and the network optimizes with free energy bound loss. Finally, we exploit the strategy of Gaussian mixture model estimation to perform the detection process with the optimized approximate distribution of each sample. The main contributions of this article can be summarized as follows:

1) To the best of our knowledge, this is the first time the normalizing flow has been adopted for anomaly detection tasks in HSI. The normalizing flow adaptively generates complex probability distributions for training samples to fit real scenarios.
2) To better enhance the discrimination between the background and anomalies, the normalizing flow is incorporated with the VAE architecture to discover the intrinsic characteristics hidden in the latent space.
3) To leverage the spatial information, we adopt the strategy of Gaussian mixture model estimation to detect the anomalies, which can better model the background statistics and increase the detection power.

The rest of this article is organized as follows. Section II describes the proposed NF-PDRD algorithm in detail. Experimental results on four data sets are conducted in Section III. Finally, Section IV concludes this article.

II. Related Works

A. VAE Architecture

VAE is a typical generative model that has been widely used in recent years. It derives from Bayesian inference and aims to solve the difficulty of computing the integral of the likelihood function with the parameter's prior distribution. In essence, the Bayesian inference problem is the acquisition of posterior probability. According to the Bayes theorem, the posterior probability can be formulated by

$$P(\mathbf{z} \mid \mathbf{x}) = \frac{P(\mathbf{z})P(\mathbf{x} \mid \mathbf{z})}{P(\mathbf{x})} \tag{1}$$

where $\mathbf{z}$ and $\mathbf{x}$ denote a single sample and latent variable, respectively. $P(\mathbf{z})$ is the prior and $P(\mathbf{x} \mid \mathbf{z})$ is the likelihood. The evidence $P(\mathbf{x})$ can be computed by

$$P(\mathbf{x}) = \int P(\mathbf{x} \mid \mathbf{z})P(\mathbf{z})dz. \tag{2}$$

However, it is intractable in most cases due to the high dimensionality of $\mathbf{z}$.

To better estimate the posterior, VAE incorporates AE with variation inference by maximizing the evidence $P(\mathbf{x})$, and we can acquire the evidence lower bound (ELBO), which is defined as follows:

$$\text{ELBO} = E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x}|\mathbf{z}\right)\right] - \mathbb{KL}\left(q_\phi\left(\mathbf{z}|\mathbf{x}\right) \| p_\theta(\mathbf{z})\right) \quad (3)$$

where ϕ and θ correspond to the encoder and decoder, respectively. $E[\cdot]$ computes the mathematical expectation and $\mathbb{KL}(\cdot)$ computes the Kullback–Leibler (KL) divergence. The first expectation part acts as the main optimization term that relates to the reconstruction error. The second KL divergence term can be regarded as a regularization term to prevent the model from collapsing, which assesses the distinction between the approximate posterior and the prior distribution. The final loss function $\mathcal{L}$ is just the opposite of ELBO, and it can unify the model's accuracy and robustness. From the perspective of probability representation, VAE extracts a unique Gaussian distribution for each sample, including the mean vector and standard deviation vector, which potentially explores the probabilistic latent space.

Compared to ordinary AE, VAE owns excellent reconstruction ability and has been successfully used in the field of anomaly detection for optical images [40]. The VAE architecture assumes the approximate posterior conforms to a multivariate Gaussian distribution. However, the approximate posterior can be any complex form in real applications. In such cases, the generative capability of VAE is limited. Unfortunately, the problem is often ignored when extracting the intrinsic characteristics in the latent space.

B. Normalizing Flows

A normalizing flow is a sequence of invertible transformations of probability distribution [41]. Considering a invertible mapping $f : \mathbb{R}^d \to \mathbb{R}^d$, the initial latent variable $\mathbf{z}$ can be changed to another variable $\mathbf{z}' = f(\mathbf{z})$ with the technique of change of variables by

$$q\left(\mathbf{z}'\right) = q(\mathbf{z})\left|\det\frac{\partial f^{-1}}{\partial \mathbf{z}'}\right| = q(\mathbf{z})\left|\det\frac{\partial f}{\partial \mathbf{z}}\right|^{-1} \quad (4)$$

where $\det(\cdot)$ is the determinant of a matrix. By concatenating several mappings, we can generate complex distributions by applying the chain rule toward (4). Through a series of K transformations f_k, the process of transformation from the initial variable $\mathbf{z}_0$ to the acquired distribution $q_K(\mathbf{z})$ is given by

$$\mathbf{z}_K = f_K \circ \ldots \circ f_2 \circ f_1\left(\mathbf{z}_0\right) \quad (5)$$

$$\log q_K\left(\mathbf{z}_K\right) = \log q_0\left(\mathbf{z}_0\right) - \sum_{k=1}^{K} \log\det\left|\frac{\partial f_k}{\partial \mathbf{z}_k}\right| \quad (6)$$

where $\circ$ represents the composition between the mapping and the variable, and the right part of (5) is the abbreviation of $f_K(f_{K-1}(\ldots f_1(x)))$. The path from the initial distribution $q(\mathbf{z}_0)$ to $q(\mathbf{z}_k)$ is called the $flow$ and the transforming path generated by the successive distributions is a $normalizing$ $flow$. With the increase of K, $q(\mathbf{z}_K)$ becomes more and more complex and the computation of $\mathbb{E}_{q_K}[g(\mathbf{z})]$is difficult. Based on the law of the unconscious statistician, it can be reformulated by

$$\mathbb{E}_{q_K}[g(\mathbf{z})] = \mathbb{E}_{q_0}\left[g\left(f_K \circ f_{K-1} \circ \ldots \circ f_1\left(\mathbf{z}_0\right)\right)\right] \quad (7)$$

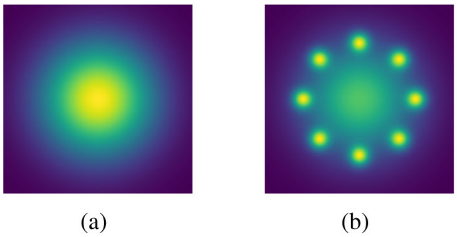


Fig. 1. Visualization of the base density as well as the transformed distribution when using normalizing flows. (a) Base distribution. (b) Transformed complex distribution.

where $g(\cdot)$ can be any invertible and smooth function.

Fig. 1 depicts the visualization of the effect with normalizing flows. Fig. 1(a) denotes the base Gaussian distribution, and Fig. 1(b) describes the complex probability distribution after the normalizing flows.

Although VAE has a powerful representation capability, the independent Gaussian distribution is not generalized for complicated scenarios to characterize the relationship between the anomalies and background pixels. With the transformation of distributions, the normalizing flow seeks to enhance the sample space and acquire actual distributions for anomalies and background pixels. As a result, discriminative features can be extracted with the normalizing flow. Since the critical point of anomaly detection lies in separating anomalies from background pixels, normalizing flow is expected to achieve good performance.

III. Methodology

A. Variational Inference With Normalizing Flows

1) Framework: Because of the powerful ability to generate various complex probability distributions, we intend to incorporate the normalizing flow into the VAE network. Fig. 2 depicts the architecture of the proposed method with five main modules: encoder $\mathbf{Enco}$, probability representation module $\mathbf{PR}$, sampling layer $\mathbf{Samp}$, normalizing flows $\mathbf{NF}$, and decoder $\mathbf{Deco}$. $\mathbf{Enco}$ consists of three fully connected layers with 400 nodes in each layer. A layer to generate the mean vector and the other layer to generate the standard deviation vector jointly constitute the $\mathbf{PR}$ module, which is regarded as the probability representation of the input data. $\mathbf{Samp}$ samples data from the standard Gaussian distribution and acquires the initial probability distribution with a simple transformation. $\mathbf{NF}$ is a normalizing flow used to transform simple distributions into complex distributions. The length of flow will be discussed in the experiment section. $\mathbf{Deco}$ consists of six fully connected layers with 400 nodes in each layer. In our experiments, we use ReLu as the activation function.

The specific training procedure is as follows: $\mathbf{Enco}$ first processes the input HSI and maps it into feature space. Then, $\mathbf{PR}$ generates the mean vector and standard deviation vector for each training sample. $\mathbf{Samp}$ combines these two vectors to produce an initial Gaussian distribution. Next, $\mathbf{NF}$ transforms the simple Gaussian distribution into complex distribution with a

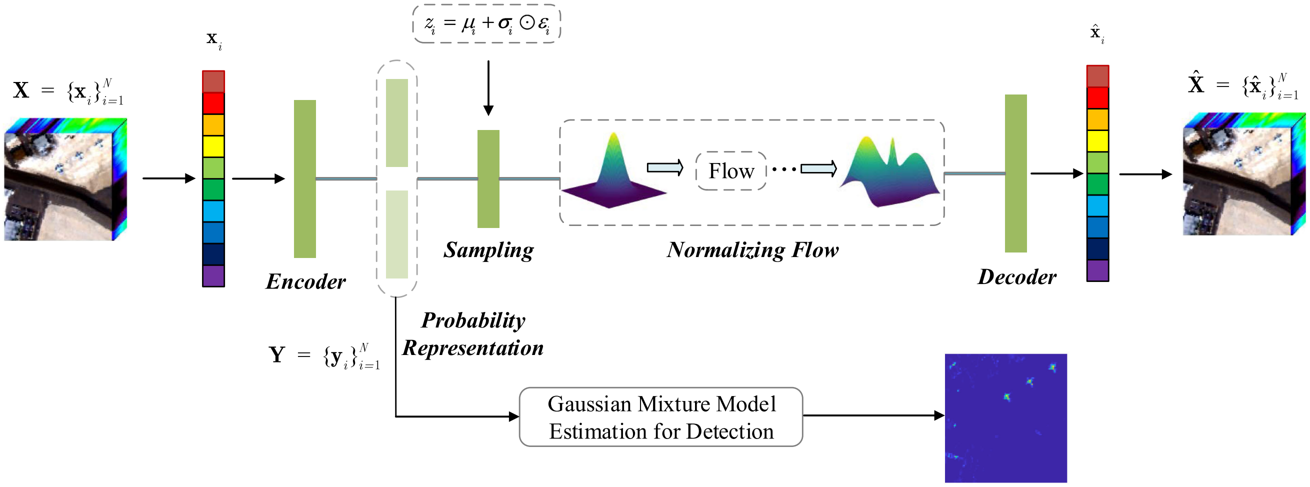

$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$
$\mathbf{x}_i$
Encoder
$z_i = \mu_i + \sigma_i \odot \varepsilon_i$
Sampling
Flow
Normalizing Flow
Decoder
$\hat{\mathbf{x}}_i$
$\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_i\}_{i=1}^N$
Probability Representation
$\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$
Gaussian Mixture Model Estimation for Detection

Fig. 2. Architecture of the proposed NF-PDRD.

sequence of flows by computing the determinant of the Jacobian matrix. Finally, **Deco** decodes the latent representation from the distribution obtained by normalizing flow and outputs the reconstructed image.

2) Loss Function of the Proposed Method: For the input HSI $\mathbf{X} = \{x_i\}_{i=1}^N \in \mathbb{R}^{B\times N}$($N$ denotes the number of samples and B denotes the number of bands), we can extract the mean vector $\boldsymbol{\mu}$ and the standard deviation vector $\boldsymbol{\sigma}$. By sampling from a normal distribution, we obtain the latent variable via $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$, which serves as the initial distribution $q_0(z_0)$. With a sequence of finite flows, the **NF** module transforms $q_0(z_0)$ into complex distribution, which is capable of characterizing the intrinsic features of the model. We evaluate the approximate posterior distribution with a sequence of flows, i.e., $q_\phi(\mathbf{z}|\mathbf{x}) := q_K(\mathbf{z}_K)$, the total loss function of our model can be expressed by

$$\begin{aligned}\mathcal{F}(\mathbf{x}) &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x}|\mathbf{z}\right)\right] + \mathbb{KL}\left(q_\phi\left(\mathbf{z}|\mathbf{x}\right)\|p_\theta(\mathbf{z})\right)\\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x},\mathbf{z})\right]\\ &= \mathbb{E}_{q_0(z_0)}\left[\log q_K\left(\mathbf{z}_K\right) - \log p\left(\mathbf{x},\mathbf{z}_K\right)\right]\\ &= \mathbb{E}_{q_0(z_0)}\left[\log q_0\left(\mathbf{z}_0\right)\right] - \mathbb{E}_{q_0(z_0)}\left[\log p\left(\mathbf{x},\mathbf{z}_K\right)\right]\\ &\quad - \mathbb{E}_{q_0(z_0)}\left[\sum_{k=1}^K \log\det\left|\frac{\partial f_k}{\partial \mathbf{z}_k}\right|\right]\\ &= \mathbb{E}_{q_0(z_0)}\left[\log q_0\left(\mathbf{z}_0\right)\right] - \mathbb{E}_{q_0(z_0)}\left[\log p\left(\mathbf{z}_K\right)\right]\\ &\quad - \mathbb{E}_{q_0(z_0)}\left[\sum_{k=1}^K \log\det\left|\frac{\partial f_k}{\partial \mathbf{z}_k}\right|\right] - \mathbb{E}_{q_0(z_0)}\left[\log p\left(\mathbf{x}|\mathbf{z}_K\right)\right]\end{aligned} \tag{8}$$

where K denotes the length of normalizing flows, and the last term $-\mathbb{E}_{q_0(z_0)}[\log p(\mathbf{x}|\mathbf{z}_K)]$ in (8) is related to the reconstruction error.

3) Invertible Finite Flow: The change of variable theorem demonstrates the fact that the transformation determines the complexity of the obtained distribution. Specifically, there are two kinds of invertible finite flows which are classified according to the mapping function: planar flows and radical flows.

Planar flows transform the variable via the mapping

$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}g\left(\mathbf{w}^\top\mathbf{z} + b\right) \tag{9}$$

where $\lambda = \{\mathbf{w} \in \mathbb{R}^d, \mathbf{u} \in \mathbb{R}^d, b \in \mathbb{R}\}$ are the parameters, d denotes the dimensionality of latent space, and $g(\cdot)$ is a smooth nonlinear mapping with derivative $g'(\cdot)$. To compute the logdet-Jacobian of $f(\mathbf{z})$, we introduce an intermediate function $\psi(\mathbf{z})$ that is characterized by

$$\psi(\mathbf{z}) = g'\left(\mathbf{w}^\top\mathbf{z} + b\right)\mathbf{w}. \tag{10}$$

Thus, the logdet-Jacobian term can be computed by

$$\det\left|\frac{\partial f}{\partial \mathbf{Z}}\right| = \left|\det\left(\mathbf{I} + \mathbf{u}\psi(\mathbf{z})^\top\right)\right| = \left|1 + \mathbf{u}^\top\psi(\mathbf{z})\right|. \tag{11}$$

Consequently, normalizing flows transform the initial distribution $q_0(z_0)$ into the complex distribution $q_K(z_K)$ by

$$\ln q_K\left(\mathbf{z}_K\right) = \ln q_0(\mathbf{z}) - \sum_{k=1}^K \ln\left|1 + \mathbf{u}_k^\top\psi_k\left(\mathbf{z}_k\right)\right|. \tag{12}$$

Furthermore, a series of transformations corresponding to radial flows can be defined as

$$f(\mathbf{z}) = \mathbf{z} + \beta g(\alpha, r)\left(\mathbf{z} - \mathbf{z}_0\right) \tag{13}$$

where $r = |\mathbf{z} - \mathbf{z}_0|, g(\alpha, r) = 1/(\alpha + r)$, $\lambda = \{\mathbf{z}_0 \in \mathbb{R}^d, \alpha \in \mathbb{R}, \beta \in \mathbb{R}\}$ are the free parameters, $\mathbf{z}_0$ is the reference point. Then, the logdet-Jacobian is computed by

$$\det\left|\frac{\partial f}{\partial \mathbf{z}}\right| = \left[1 + \beta g(\alpha, r)\right]^{d-1}\left[1 + \beta g(\alpha, r) + g'(\alpha, r)r)\right]. \tag{14}$$

Theoretically, different flows yield different mathematical explanations. By comparing (13) with (9), we observe that planar flows impose several contractions and expansions in the direction perpendicular to the hyperplane $\mathbf{w}^\top\mathbf{z} + b = 0$, while radial flows transform the latent space around the reference point. Therefore, the choice of normalizing flows can lead to totally different transformations with different properties. In real detection applications, we should first evaluate the complexity of target components and select the appropriate kind of flow.

B. Detection With Gaussian Mixture Model Estimation

By virtue of normalizing flows, we train a probability representation model with powerful generative and expressive capability. Consequently, the latent characteristics can better reflect the intrinsic properties of the original data. Thus, we seek to use the latent characteristics to accomplish the detection task.

As each pixel owns a unique Gaussian probability distribution, the background pixel can approximately be characterized by a Gaussian mixture model of its local neighborhood due to the sparseness of the anomalies, whereas anomalies cannot. Therefore, the background pixel $\mathbf{t}_b$ can be modeled by

$$p(\mathbf{t}_b) = \sum_{l=1}^{L} \pi_l \mathcal{N}\left(\mathbf{t}_{\boldsymbol{b}} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)$$

$$\text{s.t} \sum_{l=1}^{L} \pi_l = 1 \tag{15}$$

where π_l denotes the mixture coefficient, L denotes the number of Gaussian components, and $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ represent the mean value and variance, respectively.

Considering the mean value of the probability representation of HSI $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{d \times N}$, for each test pixel $\mathbf{t} \in \mathbb{R}^{d}$, we try to estimate the intensity from its local neighborhood. To eliminate the effect of the distribution of the anomalies, we employ the dual-window strategy with outer window size of $w_{\text{out}} \times w_{\text{out}}$ and inner window size of $w_{\text{in}} \times w_{\text{in}}$ to select background Gaussian components. Thus, the adjacent area $\mathbf{Y}_s = \{\mathbf{y}_i\}_{i=1}^{L} \in \mathbb{R}^{d \times L}$ is constructed for the following detection process (l denotes the number of chosen pixels; $L = w_{\text{out}} \times w_{\text{out}} - w_{\text{in}} \times w_{\text{in}}$). As $\mathbf{Y}_s$ is adaptive for each test pixel, we attempt to discover a excellent represent for $\mathbf{t}$ by

$$\arg \min_{\boldsymbol{\pi}} \|\mathbf{t} - \mathbf{Y}_s \boldsymbol{\pi}\|_2^2$$

$$\text{s.t} \sum_{l=1}^{L} \pi_l = 1 \tag{16}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_L]^{\mathrm{T}}$ is mixture coefficient vector. Notably, the term $\|\mathbf{t} - \mathbf{Y}_s \boldsymbol{\pi}\|_2^2$ estimates the reconstruction error between the test pixel $\mathbf{t}$ and its local surroundings. When $\mathbf{t}$ belongs to the background, it can be characterized by a Gaussian mixture model with its local Gaussian components. Therefore, the term $\|\mathbf{t} - \mathbf{Y}_s \boldsymbol{\pi}\|_2^2$ approaches 0. On the contrary, when $\mathbf{t}$ is a background pixel, $\|\mathbf{t} - \mathbf{Y}_s \boldsymbol{\pi}\|_2^2$ can yield a large value. In this way, the discrimination between the anomalies and the background pixels are underlined to some extent.

To acquire an expected mixture coefficient, we insert a Tikhonov regularization term to the optimization goal, which is shown as follows:

$$\arg \min_{\hat{\boldsymbol{\pi}}} \left\|\widetilde{\mathbf{t}} - \widetilde{\mathbf{Y}}_s \hat{\boldsymbol{\pi}}\right\|_2^2 + \lambda \left\|\boldsymbol{\Gamma}_{\mathbf{t}} \hat{\boldsymbol{\pi}}\right\|_2^2 \tag{17}$$

where $\widetilde{\boldsymbol{t}} = [\boldsymbol{t}; 1]$, $\widetilde{\mathbf{Y}}_s = [\mathbf{Y}_s; \mathbf{1}]$, $\mathbf{1}$ is a $1 \times L$ unit vector, and λ is a tradeoff parameter. The added dimension is to solve the sum-to-one constraint on $\boldsymbol{\pi}$. $\boldsymbol{\Gamma}_{\mathbf{t}}$ is used to adjust the weight according to the distinction between the test pixel and each pixel in the neighborhood, which is represented by

$$\boldsymbol{\Gamma}_{\mathbf{t}} = \begin{bmatrix} \|\boldsymbol{t} - \boldsymbol{y}_1\|_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \|\boldsymbol{t} - \boldsymbol{y}_L\|_2 \end{bmatrix}. \tag{18}$$

Algorithm 1: Framework of the Proposed Algorithm.

```
Input: hyperspectral data set $\mathbf{X} \in \mathbb{R}^{B \times N}$, window size
  $(w_{\text{out}}, w_{\text{in}})$, length of flow $r$, dimensionality of latent
  space $d$, and the category of flow.
Output: Anomaly detection map.
Construct a appropriate normalizing flow by (5) and
  (6);
Train a VAE architecture with normalizing flow by (8)
  and acquire mean value $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{d \times N}$ of the
  latent representation for the original data;
Compute the mixture coefficient vector $\hat{\boldsymbol{\pi}}$ through
  Gaussian mixture model estimation by (17) and (19);
Compute the anomalous degree for each pixel by (20)
  and obtain the final detection map.
```

Intuitively, if the distribution of the test pixel $\mathbf{t}$ is similar to a Gaussian component $\mathbf{y}_i$ in its neighborhood, the distance $\|\boldsymbol{t} - \boldsymbol{y}_i\|_2$ will be very small. Therefore, the corresponding mixture coefficient $\widetilde{\boldsymbol{\pi}}_i$ in (17) is allowed to hold a large value as expected. Taking derivative with regard to $\widetilde{\boldsymbol{\pi}}$ and (17) has a analytical solution:

$$\hat{\boldsymbol{\pi}} = \left(\widetilde{\mathbf{Y}}_s^T \widetilde{\mathbf{Y}}_s + \lambda \boldsymbol{\Gamma}_{\mathbf{t}}^T \boldsymbol{\Gamma}_{\mathbf{t}}\right)^{-1} \widetilde{\mathbf{Y}}_s^T \widetilde{\mathbf{t}}. \tag{19}$$

Therefore, the anomalous degree of the test pixel $\boldsymbol{t}$ can be characterized by

$$\eta = \|\mathbf{t} - \mathbf{Y}_s \hat{\boldsymbol{\pi}}\|_2 \,. \tag{20}$$

When η is larger than a threshold, $\boldsymbol{y}$ can be considered an anomaly. The overall procedure of the proposed method is shown in Algorithm 1.

IV. Experimental Results

To verify the effectiveness and superiority of the proposed method, we conducted our experiments on two real data sets and two synthetic data sets.

A. Hyperspectral Data Sets

In this section, we introduce the data sets we used for the experiments. Figs. 3 and 4 depict the pseudocolor image and ground truth maps of different data sets, respectively.

1) Real Data Sets: The San Diego data set is widely used for hyperspectral anomaly detection, which is acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the San Diego airport area, CA, USA. The spectral resolution is 10 nm, along with the spatial resolution of 20 m. With the removal of water absorption, low signal-to-noise ratio, and bad quality bands (1–6, 33–35, 94–97, 107–113, 153–166, and 221–224), we employ a subimage with the size of $100 \times 100 \times 189$

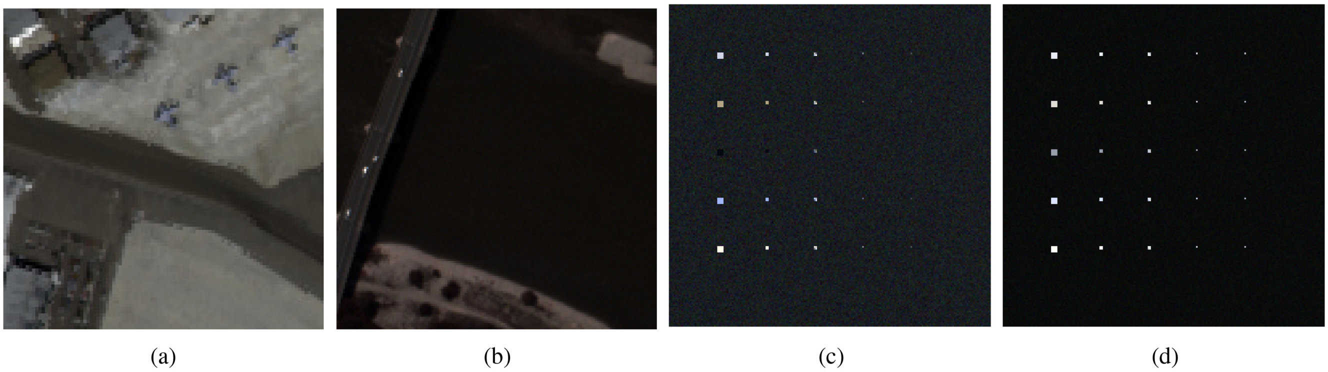


Fig. 3. Pseudocolor image of different data sets. (a) San Diego. (b) Pavia City. (c) TI. (d) TE.

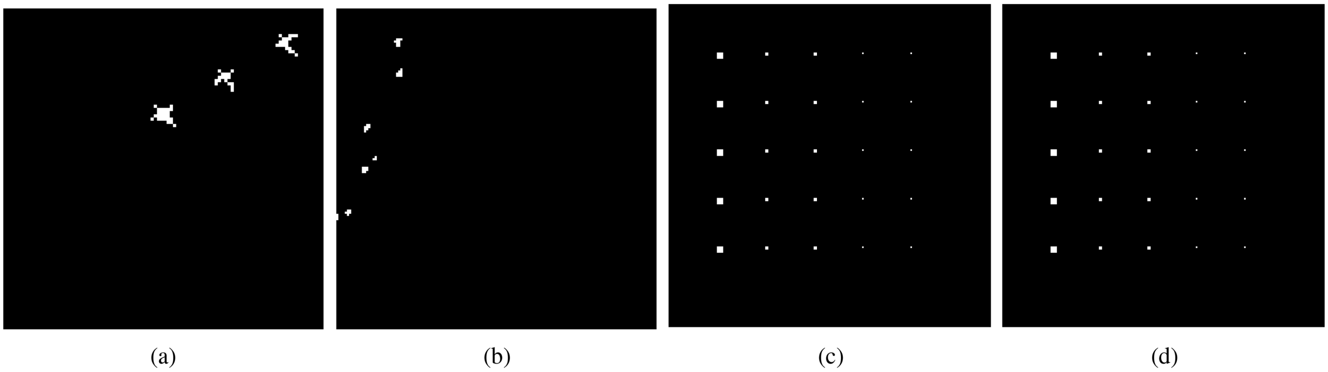


Fig. 4. Ground truth maps of different data sets. (a) San Diego. (b) Pavia City. (c) TI. (d) TE.

for the experiments. In the scene, three airplanes are considered anomalous targets. The image mainly consists of soil, parking aprons, and hangars.

The Pavia City data set is captured by the reflective optics system imaging spectrometer over the city center of Pavia in northern Italy, with a wavelength range of 430–860 nm, 205 spectral bands, and spatial size of 150×150. The ground resolution of each pixel is 1.3 m. The materials in the main background scene are bridge, water, and shadow.

2) Synthetic Data Sets: The image scene to simulate the synthetic images was collected by the AVIRIS over the Cuprite mining site, Nevada, in 1997. The data set contains 200×200 pixels and 189 spectral bands with the elimination of bad bands (1–3, 105–115, and 150–170). A total of 100 pixels that are regarded as anomalies are simulated in the data by five mineral endmembers [42]. Furthermore, two types of target insertion are designed to generate the data set.

Target implantation (TI): The background pixels are removed and replaced with clean target panels.

Target embeddedness (TE): The target panels are embedded into the clean background pixels with additive Gaussian noise.

B. Competitors

To thoroughly estimate the performance of the proposed method, we select six state-of-the-art hyperspectral anomaly methods for comparison.

1) RX [11] is a benchmark detector that employs the local statistics of the image with dual sliding windows, which assumes the background pixels obey a multivariate Gaussian distribution. The Mahalanobis distance between the test pixel and its adjacent pixels in the sliding windows is computed to estimate the anomalous degree.
2) LRASR [22] implements a matrix decomposition process and obtains a low-rank component corresponding to background pixels and a sparse component corresponding to anomalies. The band average map of the sparse component is considered as the detection result.
3) LSMAD [24] decomposes the HSI into a background component, an anomaly component, and a noise component. Then, the Mahalanobis distance is computed using the low-rank component to acquire the detection map.
4) LSDM-MoG [25] integrates a mixture of Gaussian distributions with matrix decomposition to construct complex priors, where the low-rank component is imposed a Gamma prior, and the sparse component is imposed a multinomial. The detection process can be solved by using VB.
5) AE [31] tries to reconstruct original data by the architecture of NN. Anomalies yield large reconstruction errors, while background pixels hold small reconstruction errors.
6) CRD [19] assumes the background pixel can be linearly represented by their adjacent pixels, while anomalies cannot. The representation residual of an anomaly

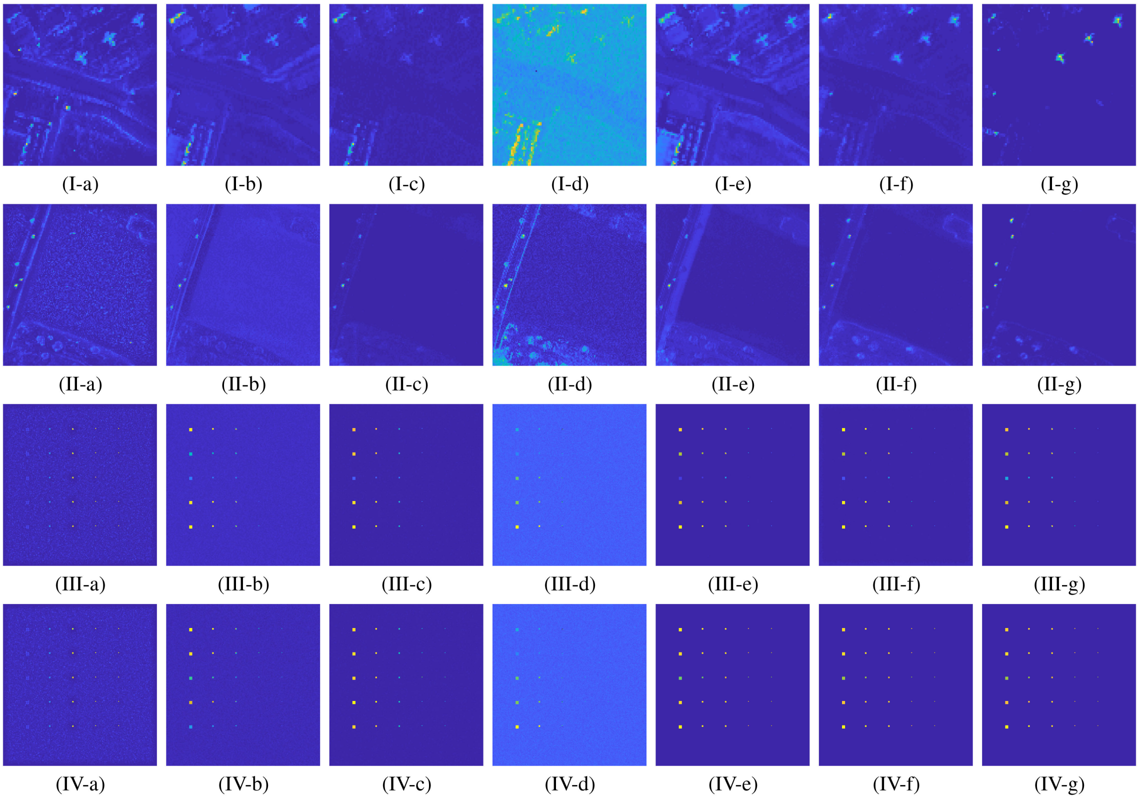


Fig. 5. Detection results of four data sets. I is the results using San Diego. II is the results using Pavia City. III is the results using TI. IV is the results using TE. (a) RX. (b) LRASR. (c) LSMAD. (d) LSDM-MoG. (e) AE. (f) CRD. (g) Proposed.

is significantly larger than a background pixel, which is used to separate the anomalies from the background pixels.

C. Evaluation Criteria

The receiver operating characteristic (ROC) curve is a widely used performance evaluation metric in hyperspectral anomaly detection. Therefore, we employ the ROC curve to estimate the detection performance of our proposed method qualitatively. For a given threshold τ, we can compute the probability of detection P_d and false alarm rate P_f according to the detection map. By setting different thresholds, we can plot continuous ROC curve with different values of P_d and P_f.

Moreover, the area under ROC curve (AUC) scores of (P_d, P_f) and (P_f, τ) are used to evaluate the detection performance quantitatively. The AUC score of (P_d, P_f) reflects the detection power of the detector, and AUC score of (P_f, τ) evaluates the performance of background suppression with different algorithms. The background-anomaly separation map are also plotted to further describe the range of intensity of anomalies and background pixels.

D. Detection Performance

1) Results of San Diego Data: The visualization of detection results on the San Diego data set is shown in Fig. 5(I). Among these methods, LSDM-MoG yields the worst detection result due to the high intensity of background pixels. We can observe that the number of detected anomalies of LRASR and LSMAD is relatively small. Thus, both LRASR and LSMAD have a low detection rate. Compared to LRASR, LSMAD, and LSDM-MoG, RX can detect more anomalies and reduce the miss detection rate to a certain degree. The background suppression AE is not adequate, as the background pixels in the bottom-left corner of the data set own high intensity. CRD can detect most anomalies, but the detection power and background suppression effect are not comparable to the proposed method.

ROC curves are displayed in Fig. 6(a) to evaluate the detection performance qualitatively. When the false alarm rate is less than 10^{-2}, the ROC curve of our proposed method stays above the other curves by a large margin, indicating the prominent detection power. Specifically, the detection probability of the proposed method is 0.2 at the beginning, while others are almost 0. When the false alarm rate is more than 10^{-2}, there are many crossovers among all methods.

AUC scores of (P_d, P_f) and (P_f, τ) are listed in Tables I and II, respectively. For the San Diego data set, the AUC score of (P_d, P_f) of the proposed method is 0.9903, which precedes other detectors to a large extent. The proposed NF-PDRD holds the lowest AUC score of (P_f, τ) among these methods, which demonstrate the powerful detection ability and excellent background suppression effect. As Fig. 7(a) displays, most of the background pixels of NF-PDRD are suppressed to a low level compared with other detectors. Moreover, almost all anomalies are brighter than the background points, suggesting an excellent separation between the anomalies and the background pixels.

2) Results of Pavia City Data: Fig. 5(II) shows the visualization of detection results on the Pavia City data set. For the

TABLE I
AUC SCORES OF (P_d, P_f) ON DIFFERENT DATA SETS

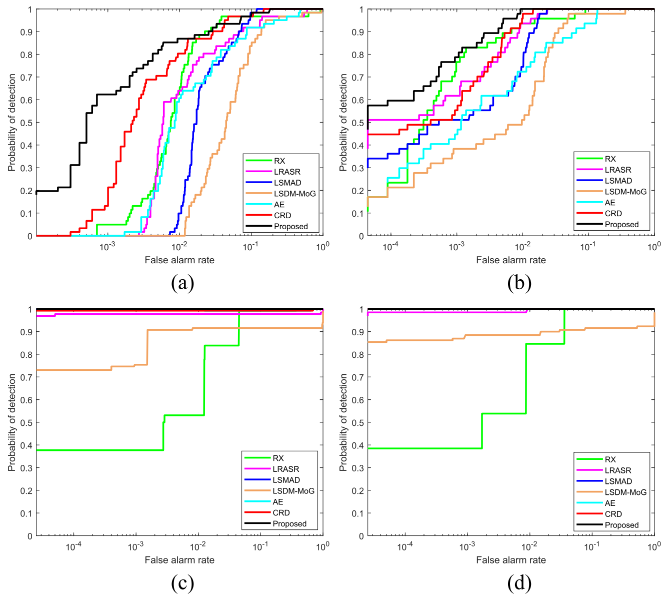
Data sets	RX	LRASR	LSMAD	LSDM-MoG	AE	CRD	Proposed
San Diego	0.9623	0.9602	0.9748	0.9208	0.9566	0.9893	**0.9903**
Pavia City	0.9955	0.9945	0.9949	0.9807	0.9849	0.9979	**0.9990**
TI	0.9860	0.9775	0.9996	0.9281	0.9951	0.9946	**1**
TE	0.9916	0.9999	1	0.9180	1	1	**1**
Average	0.9839	0.9832	0.9923	0.9369	0.9835	0.9955	**0.9973**

The bold values indicate the best performance of the corresponding method on a particular dataset.

TABLE II
AUC SCORES OF (P_f, τ) ON DIFFERENT DATA SETS

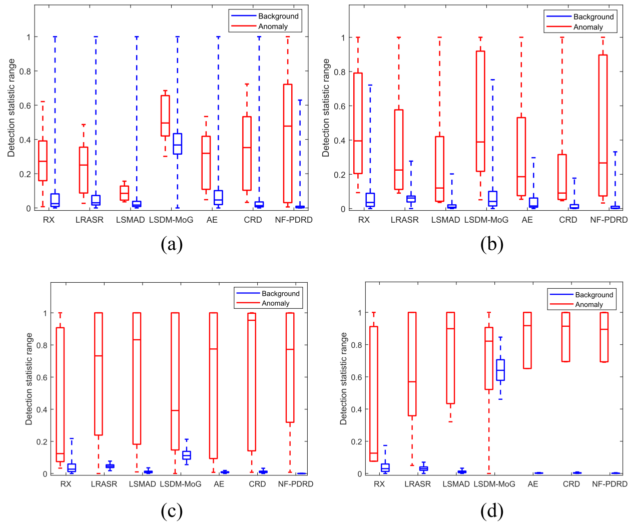
Data sets	RX	LRASR	LSMAD	LSDM-MoG	AE	CRD	Proposed
San Diego	0.0394	0.0419	0.0107	0.0464	0.0598	0.0377	**0.0023**
Pavia City	0.0177	0.0600	0.0087	0.0573	0.0272	0.0270	**0.0038**
TI	0.0321	0.1451	0.0010	0.0277	0.0009	0.0316	**0.0001**
TE	0.0331	0.1397	0.0015	0.1823	0.0024	0.0100	**0.0013**
Average	0.0306	0.0967	0.0055	0.0784	0.0226	0.0190	**0.0020**

The bold values indicate the best performance of the corresponding method on a particular dataset.



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

Fig. 6. ROC curves of the four data sets. (a) San Diego. (b) Pavia City. (c) TI. (d) TE.



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

Fig. 7. Background-anomaly separation map for different data sets. (a) San Diego. (b) Pavia City. (c) TI. (d) TE.

LSDM-MoG detector, there are many false alarms in the bottom-left of the image, causing a bad detection performance. The total intensity of RX is high, leading to a large false alarm rate. LRASR, LSMAD, AE, and CRD can achieve better detection results than RX and LSDM-MoG, but the intensity of anomalies is not significant. Our proposed method can separate anomalies from a large number of background pixels. Most anomalies preserve high intensity, and the discrimination between anomalies and background pixels is considerable.

To make a further comparison, we plot the ROC curves for the Pavia City data set as shown in Fig. 6(b). Compared to other methods, the ROC curve of the proposed method always lies in the upper-left corner of the image, which indicates the superiority of the proposed method. At the beginning of the ROC curve, the proposed NF-PDRD yields a detection rate over 0.5, which is significantly larger than other detectors.

The AUC score of (P_d, P_f) is 0.9990, which is also the largest among these methods. Furthermore, the AUC score of (P_f, τ) obtained by the proposed method owns the value 0.0038, which is close to the ideal value 0. It can be observed from Fig. 7(b) that many anomalies yield high intensity and most background pixels hold low values. Due to this fact, the background suppression effect of NF-PDRD is satisfactory on this data set.

3) Results of TI Data: For the TI data set, Fig. 5(III) depicts the visual detection results. The overall background suppression effect of LSDM-MoG is bad since most of the background pixels have high values. RX misses several anomalies, and the detection rate is rather low. The proposed method is slightly better than LRASR, LSMAD, AE, and CRD, as we can see that the anomalies located in the central row are more significant, and the background pixels are well suppressed.

TABLE III
OPTIMAL PARAMETERS OF DIFFERENT DETECTORS ON FOUR DATA SETS

Data sets	RX	LSMAD	LSDM-MoG	CRD
San Diego	$w_{\text{out}} = 19, w_{\text{in}} = 9$	$r = 5$	$l_0 = 10, K = 4$	$w_{\text{out}} = 27, w_{\text{in}} = 11$
Pavia City	$w_{\text{out}} = 17, w_{\text{in}} = 11$	$r = 4$	$l_0 = 12, K = 4$	$w_{\text{out}} = 15, w_{\text{in}} = 9$
TI	$w_{\text{out}} = 17, w_{\text{in}} = 9$	$r = 4$	$l_0 = 6, K = 4$	$w_{\text{out}} = 17, w_{\text{in}} = 7$
TE	$w_{\text{out}} = 17, w_{\text{in}} = 9$	$r = 5$	$l_0 = 12, K = 4$	$w_{\text{out}} = 15, w_{\text{in}} = 7$

Fig. 6(c) depicts the ROC curves of these algorithms. The RX and LSDM-MoG stay below other methods, and they only intersect with other detectors when the detection probability is 1. The ROC curves of other methods are similar; so it is necessary to compute the AUC scores to evaluate the performance of our proposed method.

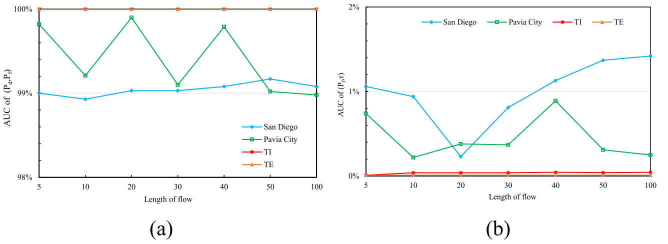
As Tables I and II display, the proposed method acquires the best AUC score of (P_d, P_f) and (P_f, τ). It is worth mentioning that the AUC score of (P_d, P_f) is 1 and (P_f, τ) is very close to 0, manifesting the effectiveness of background suppression and powerful detection ability. Furthermore, we can know from Fig. 7(c) that nearly all anomalies hold larger values than background pixels for NF-PDRD. Therefore, the discrimination between the anomalies and the background pixels is evident.

4) Results of TE Data: Fig. 5(IV) illustrates the visualization of detection results for the TE data set. More than half of the anomalies are not detected in the results of the LSDM-MoG detector since the imposed complex prior distribution may not make a difference. As we can see, RX, LRASR, and LSMAD fail to detect a certain number of anomalies. Therefore, the detection performance is not excellent to some extent. The detection map of the proposed method is similar to that of AE and CRD because all anomalies can be detected in these algorithms.

ROC curves for the TE data are displayed as shown in Fig. 6(d). Notably, except for RX, LSDM-MoG, and LRASR, the ROC curves of several methods overlap, causing it hard to further discriminate the differences.

As for the AUC score of (P_d, P_f), LSDM-MoG is the lowest among these methods. The distinctions between the proposed NF-PDRD and other methods are trivial, as the values are nearly 1. Moreover, the proposed method achieves the lowest AUC score of (P_f, τ), which verifies the background suppression ability. Fig. 7(d) demonstrates that the anomaly with the smallest value is brighter than the background pixel with the largest value by a large margin, indicating a complete separation between the anomalies and the background pixels and the anomalies. Thus, the NF-PDRD achieves excellent detection performance.

The NF-PDRD and low rank based algorithms such as LRASR yield discriminative and representative powers. LRASR transforms the data from image space to background space with distinctive features. The selection of the background dictionary is critical to the detection performance, and it remains optimized to adapt to real situations continuously. NF-PDRD tries to generate expressive latent representation with the consideration of original distribution. With the incorporation of normalizing flows, the decoder generates the reconstructed sample from a more generalized distribution, which enhances the sample space and makes the difference between the anomalies and background more discriminative.



(a) (b)

Fig. 8. Parametric analysis of the length of flow. (a) AUC of (P_d, P_f). (b) AUC of (P_f, τ).

E. Parametric Analysis

To acquire the optimal parameters of these detectors, we conduct extensive experiments with the guidance of corresponding parameter selecting strategies. Table III shows the optimal parameters of these methods. Specifically, w_{out} and w_{in} represent the outer and inner window sizes of RX and CRD. β and λ denote the tradeoff terms of low-rankness and sparsity of the LRASR algorithm. r indicates the low-rank degree of LSMAD. l_0 and K signify the initial rank and the number of mixture Gaussian noise, respectively. Furthermore, the parameters' impacts on the experimental results, including the length of flow, the dimensionality of latent space, and the category of flow, are analyzed comprehensively.

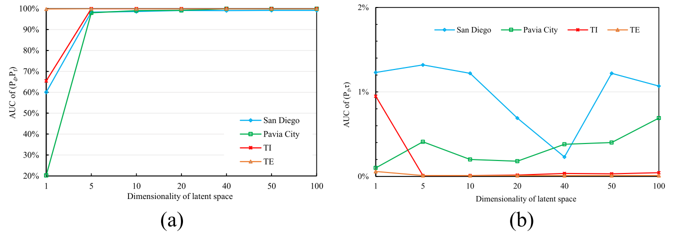
The length of flow r controls the scale of architecture. It is a vital parameter concerning the generative capability of the model. If it is too large, we will suffer expensive computational costs. While it is too small, the effectiveness of the model cannot be guaranteed. With thorough consideration, we set the value range of this parameter to [5, 10, 20, 30, 40, 50, 100]. The AUC scores of (P_d, P_f) and (P_f, τ) with the change of r are depicted in Fig. 8. Since the curves are not monotonically increasing or decreasing, there is no explicit principle of how the parameter r affects the AUC scores. On the other hand, the computational cost will increase as r rises. Therefore, the optimal r are 20, 10, 10, and 10 for San Diego, Pavia City, TI, and TE data sets, respectively.

The dimensionality of latent space d is a crucial parameter for the proposed method, which controls the reconstruction ability and the separability of anomalies and background pixels. The range of d we set in the experiment is [1, 5, 10, 20, 40, 50, 100]. As we can observe from Fig. 9, although the optimal parameters of d are different for different data sets, the proposed detector can achieve good performance when d is set to an intermediate value. From the perspective of interpretability, the phenomenon is reasonable as a proper value of d guarantees effective probabilistic representation for the intrinsic structure

TABLE IV
Execution Time (in Seconds) for Different Data Sets

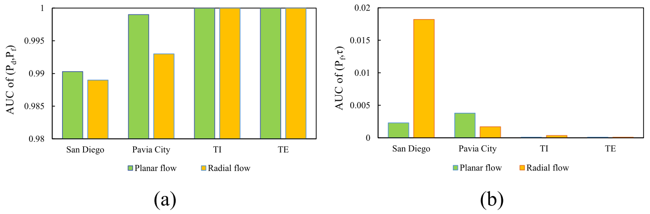
Data sets	RX	LRASR	LSMAD	LSDM-MoG	AE	CRD	Proposed
San Diego	40.06	24.33	**8.62**	11.78	51.28	571.62	44.32
Pavia City	54.72	139.21	**10.83**	39.42	56.11	71.80	54.25
TI	182.74	57.61	46.63	19.10	**3.67**	329.16	295.94
TE	173.56	136.31	36.64	41.64	**7.11**	172.60	310.82
Average	112.77	89.37	**25.68**	27.99	29.54	286.30	176.33

The bold values indicate the best performance of the corresponding method on a particular dataset.



(a) (b)

Fig. 9. Parametric analysis of the dimensionality of latent space. (a) AUC of (P_d, P_f). (b) AUC of (P_f, τ).



(a) (b)

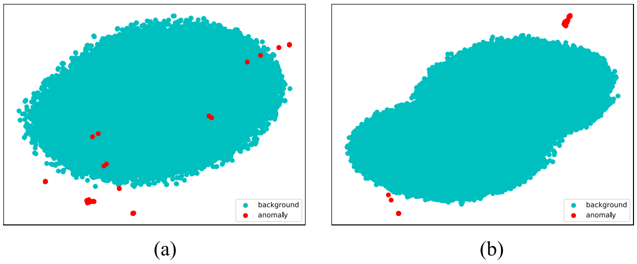
Fig. 10. Parametric analysis of the category of flow. (a) AUC of (P_d, P_f). (b) AUC of (P_f, τ).

of the original data. Specifically, the optimal values of d are 40, 20, 20, and 20 for San Diego, Pavia City, TI, and TE, respectively.

The category of flow determines how we formulate the network. As we know, normalizing flows include planar flows and radical flows. Thus, we evaluate the performance of these two kinds of flows. As shown in Fig. 10, we can get higher AUC score of (P_d, P_f) and lower AUC score of (P_f, τ) with the use of planar flow. Theoretically, radical flow can generate more complicated distributions than planar flow with more computational time. The prior distribution of these four scenes is not so complex. Therefore, we exploit the planar flow for the four data sets.

F. Execution Time

The execution time of the above detectors on the four data sets are listed in Table IV. All experiments were performed on a computer with a 64-b Intel i7-8700 CPU of 3.2 GHz on Windows 10. In total, local methods take much more time than matrix decomposition based methods. Specifically, CRD has the highest execution time among these methods. Our proposed algorithm is slightly faster than CRD due to the probability distribution representation.



(a) (b)

Fig. 11. 2-D visualization of the distribution of anomalies and background of the latent features on TI data set. (a) NF-PDRD without NF. (b) NF-PDRD.

TABLE V
AUC Scores of (P_d, P_f) With Different Components or Detectors on Four Data Sets

	San Diego	Pavia City	TI	TE
PDRD	0.9847	**0.9993**	0.9847	**1**
VAE with RX	0.9426	0.9658	0.9125	0.9759
NF-PDRD without NF	0.9814	0.9962	0.9473	**1**
NF-PDRD	**0.9903**	0.9990	**1**	**1**

The bold values indicate the best performance of the corresponding method on a particular dataset.

TABLE VI
AUC Scores of (P_f, τ) With Different Components or Detectors on Four Data Sets

	San Diego	Pavia City	TI	TE
PDRD	0.0606	0.0213	0.0874	0.0165
VAE with RX	0.0607	0.0480	0.0973	0.0345
NF-PDRD without NF	0.0203	0.0694	0.0983	0.0053
NF-PDRD	**0.0023**	**0.0038**	**0.0001**	**0.0013**

G. Ablation Study

To verify the effectiveness of normalizing flow, we conduct component analysis through extensive experiments with visual and quantitative comparisons. We keep the VAE architecture and the Gaussian mixture model estimation detector identical with the elimination of the normalizing flow component, and it is termed NF-PDRD without NF. Fig. 11 shows the distribution of anomalies and background in the latent space for the TI data set with the technique of t-SNE [43]. The anomalies are separated from the background pixels with the normalizing flow, and the discrimination between the anomalies and background pixels is significantly enhanced. Moreover, the quantitative comparisons are listed in Tables V and VI. The AUC scores of both (P_d, P_f) and (P_f, τ) illustrate that the normalizing flow can increase the detection power and background suppression effect.

Furthermore, we compare the proposed method with the other two VAE-based detectors. To make a fair comparison, the VAE architecture of these methods yield identical network and parameters. The difference lies in the anomaly detector. PDRD [44] adopts a Wasserstein distance based detector, while VAE with RX employs a Mahalanobis distance based anomaly detector. The experimental results manifest that NF-PDRD has the best background suppression effect among these methods. The (P_f, τ) of the proposed method on four data sets are all near 0. The detection power of NF-PDRD is also excellent, and we can see the (P_d, P_f) of the proposed method leads other detectors by a large margin on San Diego and TI. From the above analysis, we can conclude that the normalizing flow can enhance the detection power and background suppression effect of the model.

V. Conclusion

It is still challenging for deep learning based models to convey a theoretical explanation of their powerful capability in the task of hyperspectral anomaly detection. In this article, we present a novel NF-PDRD framework for hyperspectral anomaly detection, which first introduce a normalizing flow into the VAE model. Subsequently, the Gaussian approximate posterior can be transformed into a much more complex distribution used to generate the output data. This process makes the whole network generative and expressive, which can simulate the various real scenarios. Finally, the strategy of Gaussian mixture model estimation is used to detect the anomalies by fully utilizing spatial knowledge. Experiments on both synthetic and real data sets verify the superiority and effectiveness of the proposed NF-PDRD algorithm. However, it also needs to be highlighted that the consuming time of the proposed method is relatively high, where there is still much room to optimize. In future works, we can introduce the local similarity constraint into the optimization process as an extra loss term and incorporate several global architectures to reduce time complexity. Therefore, the local detector can be replaced by a global detector without loss of spatial characteristics. Moreover, we intend to add more generative models and background suppression techniques to the framework.

References

[1] Z. Li and Y. Zhang, "Hyperspectral anomaly detection via image super-resolution processing and spatial correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2307–2320, Mar. 2021.

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

[3] S. Song, H. Zhou, Y. Yang, and J. Song, "Hyperspectral anomaly detection via convolutional neural network and low rank with density-based clustering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3637–3649, Sep. 2019.

[4] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 5–28, Jul. 2010.

[5] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proc. IEEE*, vol. 97, no. 6, pp. 1031–1055, Jun. 2009.

[6] S. Khazai, A. Safari, B. Mojaradi, and S. Homayouni, "An approach for subpixel anomaly detection in hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 769–778, Apr. 2013.

[7] H. Su, Z. Wu, Q. Du, and P. Du, "Hyperspectral anomaly detection using collaborative representation with outlier removal," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 5029–5038, Dec. 2018.

[8] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[9] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3318–3329, Jul. 2020.

[10] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Joint reconstruction and anomaly detection from compressive hyperspectral images using Mahalanobis distance-regularized tensor RPCA," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2919–2930, May 2018.

[11] I. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[12] Q. Ling, Y. Guo, Z. Lin, and W. An, "A constrained sparse representation model for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2358–2371, Apr. 2019.

[13] Q. Guo, B. Zhang, Q. Ran, L. Gao, J. Li, and A. Plaza, "Weighted-RXD and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2351–2366, Jun. 2014.

[14] M. Carlotto, "A cluster-based approach for detecting man-made objects and changes in imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 374–387, Feb. 2005.

[15] H. Kwon, S. Z. Der, and N. M. Nasrabadi, "Adaptive anomaly detection using subspace separation for hyperspectral imagery," *Opt. Eng.*, vol. 42, no. 11, pp. 3342–3351, 2003.

[16] H. Kwon and N. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388–397, Feb. 2005.

[17] J. Zhou, C. Kwan, B. Ayhan, and M. T. Eismann, "A novel cluster kernel RX algorithm for anomaly and change detection using hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6497–6504, Nov. 2016.

[18] L. Zhu, G. Wen, and S. Qiu, "Low-rank and sparse matrix decomposition with cluster weighting for hyperspectral anomaly detection," *Remote Sens.*, vol. 10, no. 5, pp. 1–17, 2018.

[19] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.

[20] W. Sun, C. Liu, J. Li, Y. M. Lai, and W. Li, "Low-rank and sparse matrix decomposition-based anomaly detection for hyperspectral imagery," *J. Appl. Remote Sens.*, vol. 8, no. 1, pp. 1–18, 2014.

[21] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[22] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 1990–2000, Apr. 2016.

[23] S. Yu, X. Li, L. Zhao, and J. Wang, "Hyperspectral anomaly detection based on low-rank representation using local outlier factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1279–1283, Jul. 2021.

[24] Y. Zhang, B. Du, L. Zhang, and S. Wang, "A low-rank and sparse matrix decomposition-based Mahalanobis distance method for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1376–1389, Mar. 2016.

[25] L. Li, W. Li, Q. Du, and R. Tao, "Low-rank and sparse decomposition with mixture of gaussian for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4363–4372, Sep. 2021.

[26] T. Cheng and B. Wang, "Total variation and sparsity regularized decomposition model with union dictionary for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1472–1486, Feb. 2021.

[27] L. Zhang and B. Cheng, "Transferred CNN based on tensor for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2115–2119, Dec. 2020.

[28] W. Li, G. Wu, and Q. Du, "Transferred deep learning for anomaly detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 597–601, May 2017.

[29] J. Lei, W. Xie, J. Yang, Y. Li, and C.-I. Chang, "Spectral–spatial feature extraction for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8131–8143, 2019.

[30] N. Ma, Y. Peng, S. Wang, and P. H. Leong, "An unsupervised deep hyperspectral anomaly detector," *Sensors*, vol. 18, no. 3, pp. 1–21, 2018.

[31] E. Bati, A. Çalışkan, A. Koz, and A. A. Alatan, "Hyperspectral anomaly detection method based on auto-encoder," *Image Signal Process. Remote Sens. XXI*, vol. 9643, pp. 220–226, 2015.

[32] B. Hosseiny and R. Shah-Hosseini, "A hyperspectral anomaly detection framework based on segmentation and convolutional neural network algorithms," *Int. J. Remote Sens.*, vol. 41, no. 18, pp. 6946–6975, 2020.

[33] J. Lei, S. Fang, W. Xie, Y. Li, and C.-I. Chang, "Discriminative reconstruction for hyperspectral anomaly detection with spectral learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7406–7417, Oct. 2020.

[34] W. Xie, B. Liu, Y. Li, J. Lei, C.-I. Chang, and G. He, "Spectral adversarial feature learning for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2352–2365, Apr. 2020.

[35] S. Arisoy, N. M. Nasrabadi, and K. Kayabol, "Unsupervised pixel-wise hyperspectral anomaly detection via autoencoding adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[36] T. Jiang, Y. Li, W. Xie, and Q. Du, "Discriminative reconstruction constrained generative adversarial network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4666–4679, Jul. 2020.

[37] K. Jiang, W. Xie, Y. Li, J. Lei, G. He, and Q. Du, "Semisupervised spectral learning with generative adversarial network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 5224–5236, Jul. 2020.

[38] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[40] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture IE*, vol. 2, no. 1, pp. 1–18, 2015.

[41] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 32nd Int. Conf. Mach. Learn.*, F. Bach and D. Blei, eds., vol. 37 of Proceedings of Machine Learning Research, Lille, France, pp. 1530–1538, 2015.

[42] C.-I. Chang, H. Cao, S. Chen, X. Shang, C. Yu, and M. Song, "Orthogonal subspace projection-based go-decomposition approach to finding low-rank and sparsity matrices for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2403–2429, Mar. 2021.

[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[44] S. Yu, X. Li, S. Chen, and L. Zhao, "Exploring the intrinsic probability distribution for hyperspectral anomaly detection," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 441.

Xiaorun Li received the B.S. degree from the National University of Defense Technology, Changsha, China, in 1992, and the M.S. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1995 and 2008, respectively, all in control theory and control engineering.

Since 1995, he has been with Zhejiang University, where he is currently a Professor with the College of Electrical Engineering. His research interests include hyperspectral image processing, signal and image processing, and pattern recognition.

Shaoqi Yu received the B.S. degree in measurement and control technology and instruments from the Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2014. He is currently working toward the Ph.D. degree in control theory and control engineering with Zhejiang University, Hangzhou, China.

His research interests included image processing, machine learning, and pattern recognition.

Shuhan Chen (Member, IEEE) received the B.S. degree from Ludong University, Yantai, China, in 2011, the M.S. degree from Liaoning Technical University, Huludao, China, in 2014, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2020, all in control theory and control engineering.

She was a visiting Faculty Research Assistant with the Remote Sensing Signal and Image Processing Laboratory (RSSIPL), Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA, from 2018 to 2020. She is currently a Postdoc with the Department of Electrical Engineering, Zhejiang University, Hangzhou, China. Her research interests include hyperspectral image processing and multisource remote sensing image registration.

Liaoying Zhao received the B.S. and M.S. degrees from Hangzhou Dianzi University, Hangzhou, China, in 1992 and 1995, respectively, and the Ph.D. degree from Zhejiang University, Hangzhou, in 2004, all in electrical engineering.

Since 1995, she has been with Hangzhou Dianzi University, where she is currently a Professor with the College of Computer Science. Her research interests include hyperspectral image processing, signal and image processing, pattern recognition, and machine learning.