E2E-LIADE: End-to-End Local Invariant Autoencoding Density Estimation Model for Anomaly Target Detection in Hyperspectral Image

Kai Jiang*, Weiying Xie†, Member, IEEE, Jie Lei*, Member, IEEE, Zan Li*, Senior Member, IEEE, Yunsong Li†, Member, IEEE, Tao Jiang*, Student Member, IEEE, and Qian Du*, Fellow, IEEE

Abstract—Hyperspectral anomaly target detection (also known as hyperspectral anomaly detection (HAD)) is a technique aiming to identify samples with atypical spectra. Although some density estimation-based methods have been developed, they may suffer from two issues: 1) separated two-stage optimization with inconsistent objective functions makes the representation learning model fail to dig out characterization customized for HAD and 2) incapability of learning a low-dimensional representation that preserves the inherent information from the original high-dimensional spectral space. To address these problems, we propose a novel end-to-end local invariant autoencoding density estimation (E2E-LIADE) model. To satisfy the assumption on the manifold, the E2E-LIADE introduces a local invariant autoencoder (LIA) to capture the intrinsic low-dimensional manifold embedded in the original space. Augmented low-dimensional representation (ALDR) can be generated by concatenating the local invariant constrained by a graph regularizer and the reconstruction error. In particular, an end-to-end (E2E) multidistance measure, including mean-squared error (MSE) and orthogonal projection divergence (OPD), is imposed on the LIA with respect to hyperspectral data. More important, E2E-LIADE simultaneously optimizes the ALDR of the LIA and a density estimation network in an E2E manner to avoid the model being trapped in a local optimum, resulting in an energy map in which each pixel represents a negative log likelihood for the spectrum. Finally, a postprocessing procedure is conducted on the energy map to suppress the background. The experimental results demonstrate that compared to the state of the art, the proposed E2E-LIADE offers more satisfactory performance.

I. INTRODUCTION

HYPERSONTICAL remote sensing, which plays a pivotal role in Earth observation and deep space exploration, collects a large number of measured wavelength bands from visible to infrared range of the electromagnetic spectrum [1]. On this basis, hyperspectral image (HSI) interpretation techniques, including classification, target detection, and anomaly target detection, have emerged and explored [2]–[9]. Among them, the hyperspectral anomaly target detection (HAD), which is unsupervised detection of targets that are spatially or spectrally different from the surrounding background without any prior information (e.g., unknown targets or military camouflage targets), has drawn increasing attention in practical situations [10].

Existing HAD methods can be grouped into two major categories: 1) traditional and 2) deep learning-based HAD. Regarded as one of the most well-known traditional methods, the Reed XiaoLi (RX) algorithm, proposed by Reed and Yu [11], is essentially based on the multivariate Gaussian assumption for background distribution. Then, an improved variant called the local RX (LRX) [12] was presented. Furthermore, a Gaussian mixture model (GMM)-based method called the cluster-based anomaly detection (CBAD) [13] showed better performance in estimating the background distribution. In [14], a robust background regression-based model aimed to implement robust background estimation and HAD. Different from the above statistical-based detector, various sparse representation-based HAD methods have been proposed, such as the collaborative-representation-based detector (CRD) [15] and background joint sparse representation (BJSR) detector [16]. Recently, based on the assumption that the background is low rank and anomalies with sparse property are preserved in the residual image, low-rank and sparse matrix decomposition (LRaSMD) has been widely discussed. Zhang et al. [17] proposed an LRaSMD-based Mahalanobis
distance method for anomaly detection (LSMAD), which fully explores the low rank prior knowledge of the background and applies the Mahalanobis distance to detect probable anomalies. In addition, Xu et al. [18] introduced a low-rank and sparse representation (LRASR) anomaly detector, which models the sparse component by constructing a dictionary and indicates anomalies with an l2-norm of the sparse component. Furthermore, Li et al. [19] constructed a detector based on LRaSMD in conjunction with a mixture noise model, which accurately describes various noise characteristics. Unlike the above-mentioned works, Yuan et al. [20] proposed a graph-based method to detect anomaly targets, and Kang et al. [21] proposed an HAD algorithm based on the attribute filter and edge-preserving filter, which can efficiently achieve competitive detection capability.

More recently, deep learning-based methods, which can construct an effective, hierarchical, abstract, and high-level representation for background, have attracted more interest. Li et al. [22] proposed a transferred deep convolutional neural network (CNN)-based strategy for HAD, where a reference dataset was employed to train the CNN with pixel pairs generated from labeled samples. In [23], a novel HAD algorithm, based on CNN and low-rank representation (LRR), was applied to extract the abundance map with CNN and developed an improved LRR method to detect anomalies. Besides that, Zhang and Cheng [24] constructed a stacked autoencoder (SAE)-based adaptive subspace model (SAEASM), in which in-depth features of differences are extracted to perform HAD. Furthermore, in order to extract more discriminative representation, Xie et al. [26] introduced a spectral constrained adversarial AE (SC_AAE) to suppress background samples while the characteristic of anomalies can be maintained. In [27], a spectral adversarial feature learning (SAFL) structure for HAD was presented to extract the intrinsic spectral features in deep latent space and better distinguish between anomalies and background. Ma et al. [28] applied a deep belief networks-based HAD algorithm (DBN-AD). Afterward, a generative adversarial net (GAN)-based framework was constructed to extract a discriminative background representation with anomalous spectra suppression [29]. The above-mentioned representation learning methods for HAD can alleviate the curse of dimensionality. However, such two-step approaches, in which representation learning is performed to conduct feature extraction first and anomaly target detection is performed on the extracted features subsequently, are separately learned [30], [31]. Since only generic fidelity loss functions rather than the customized objective for anomaly target detection are employed, the representation learning model is trained without the guidance from the subsequent detection procedure. Consequently, during the dimensionality reduction procedure, the characterization customized for HAD may be lost, and the two-step approach may yield a suboptimal solution. Therefore, there is an urgent need to explore a new framework to jointly optimize the feature extractor and the estimation network to provide better performance gains in HAD in an unsupervised fashion.

To approximate the optimal solution while satisfying the assumption on manifold, we propose a unified framework, called the end-to-end local invariant autoencoding density estimation (E2E-LIADE), for the first time in the field of HAD. With regard to end-to-end (E2E), it means that each input spectral vector directly corresponds to an anomalous density estimation, which differs from the aforementioned two-step framework. The proposed E2E-LIADE is based on two assumptions: 1) anomalies always reside in low-density area and 2) anomalies can be distinguished from the background not only in a reduced low-dimensional representation space but also in the reconstruction space. The low-density assumption of anomalies are derived from the statistical-based methods, such as multivariate Gaussian models [11], GMM [13] and...
so on [32], [33]. Learning with an alternating algorithm, that is, the expectation–maximization (EM) algorithm [34], GMM can hardly perform joint optimization for representation learning and density estimation, which leads to the necessity of adopting a suboptimal two-step approach. Considering that GMM or other statistical-based methods are always stuck in suboptimal solutions due to the curse of dimensionality when the number of spectral bands increases [30], the proposed E2E-LIADE directly develops an estimation network and incorporates it into a deep feature extractor to jointly learn the parameters. Moreover, based on the assumption on manifold, a spectrum significantly different from its neighbors is regarded to be anomalous. Thus, the local structural information of hyperspectral data constitutes a more rational discriminative representation for differentiating anomalies from the background in both a low-dimensional representation space and its reconstruction space. According to the aforementioned analysis, the proposed E2E-LIADE overcomes the problem that local structural information of hyperspectral data cannot be characterized by ordinary representation learning models such as AE, which are constrained by a generic fidelity loss function.

To begin with, this article proposed a local invariant AE (LIA) to capture an intrinsic low-dimensional manifold, which is embedded in the original high-dimensional space for satisfying the assumption on manifold. The local invariant low-dimensional representation solved by a graph regularizer and the spectral reconstruction error obtained from the AE are concatenated as augmented low-dimensional representation (ALDR), which is utilized to alleviate the curse of dimensionality and fully explore the local structural information of hyperspectral dataset. Particularly, mean-squared error (MSE) and orthogonal projection divergence (OPD) are incorporated as an E2E multidistance measure, which is imposed on the LIA with respect to hyperspectral data. In addition, an estimation network is proposed and incorporated into the feature extractor to estimate the distribution of the ALDR with complex structures. We use the mixture membership to update the estimation network parameters, and the probability distribution estimated by the network is utilized to perform HAD. As a result, an energy map in which each pixel refers to a negative log likelihood for the spectrum can be generated. By jointly optimizing the LIA and the estimation network with stochastic gradient descent (SGD) [35] in an E2E manner, the ALDR, which directly enhances the anomaly target detection task, can be learned with a trainable objective customized for HAD. Furthermore, in light of the hypothesis that targets with distinguished intensity and tiny size properties in the spatial domain always exhibit a high likelihood of being anomalous targets, a postprocessing procedure is proposed to perform background suppression and further decrease false alarms.

Compared with other HAD frameworks, the major contribution of our framework can be summarized as follows.

1) For the first time, a novel E2E-LIADE framework, which simultaneously optimizes the feature extractor and the estimation network in an E2E manner, is proposed to effectively preserve the low-dimensional representation customized for HAD, avoid the model being stuck in suboptimal solutions, and adapt to variant scenes more easily.

2) To satisfy the assumption on manifold, an LIA, which captures the intrinsic low-dimensional manifold embedded in the original high-dimensional space by a graph regularizer and the spectral reconstruction error, is introduced to fully utilize the local structural information of hyperspectral dataset for HAD.

3) Specifically designed for hyperspectral data, a novel E2E multidistance measure is embedded into the LIA in order to completely consider the spatial distance and spectral distance. In particular, OPD is suitable for calculating the differences between two spectral vectors theoretically and experimentally.

4) With the purpose of achieving comprehensive estimation, the ALDR concatenates the local invariant low-dimensional representation as well as the reconstruction error, and jointly trains with the proposed estimation network, which creates an energy map in which the spectra with low density are most likely anomalies.

5) The experiments conducted on a series of datasets captured by various sensors demonstrate that the proposed method has superiority in terms of detection accuracy and generalization capability compared with the state-of-the-art methods.

The remainder of this article is organized as follows. Section II gives the detailed description on the proposed HAD framework. Section III presents the experimental results and analysis. Finally, the conclusions are drawn in Section IV.

II. PROPOSED HAD FRAMEWORK

As mentioned earlier, although existing two-step HAD approaches based on feature extraction, followed by anomaly target detection, have successfully reduced the dimensionality of the data, a weakness of these two-step approaches is the lack of the customized objective for HAD during dimensionality reduction. Consequently, the characterization customized for HAD cannot be effectively dug out, and the two-step approaches may generate a suboptimal solution. In order to address the aforementioned issue, a unified E2E-LIADE framework, which jointly performs dimensionality reduction and distribution density estimation in an E2E manner, is proposed. The proposed E2E-LIADE focuses on generating the low-dimensional representation customized for HAD and preserving the inherent information from the original high-dimensional spectral space. Fig. 1 indicates the schematic diagram of our proposed E2E-LIADE.

Let an original hyperspectral image with $M \times N$ pixels and $B$ spectral bands be denoted as $X \in \mathbb{R}^{M \times N \times B}$. Let $x_i \in \mathbb{R}^{B \times 1}, (i = 1, 2, \ldots, L, L = M \times N)$ represent the $i$th spectral vector sampled from the dataset $X$. Then, the original hyperspectral dataset can be reshaped as

$$X = [x_1, x_2, \ldots, x_L] \in \mathbb{R}^{L \times B}. \quad (1)$$

Density estimation-based methods assume that a normal instance is generated by a distribution with parameters $\Theta$ and probability density function (pdf) $f(x; \Theta), x_i \in X$, while...
anomalies always reside in a low-density area. Thus, the inverse of the pdf \( f(x; \Theta) \) can be adopted as an anomaly score for the instance \( x_i \). The parameters \( \Theta \) can be estimated via maximum-likelihood estimation, which can be formulated as the following constrained optimization problem:

\[
\max_{\Theta} \sum_{i=1}^{L} \ln(f(x_i; \Theta)) \\
\text{s.t. } 0 \leq f(x_i; \Theta) \leq 1.
\]  

(2)

In a real application, the distribution of hyperspectral background is complicated and does not conform with multivariate Gaussian distribution. A GMM is a weighted finite sum of Gaussian components as shown in Fig. 2. Based on the law of large numbers and the strong representational capability of GMM, it is used to estimate distribution for HAD. The pdf of GMM in the multivariate occasion can be formulated as:

\[
f(x; \Theta) = \sum_{k=1}^{K} \phi_k N(x_i; \mu_k, \Sigma_k) \\
\Theta = \{\phi, \mu, \Sigma\}
\]  

(3)

where \( \mu_k \) refers to the mean vector, the non-negative weight \( \phi_k \) sums to 1, and the covariance matrix \( \Sigma_k \) is positive definite. A variety of approaches to the problem of GMM parameter estimation has been proposed, many of which focus on maximum-likelihood estimation, such as EM, gradient descent (GD), and so on. To estimate the parameters of GMM, we are interested in maximizing the constrained optimization problem as follows:

\[
\max_{\phi, \mu, \Sigma} \sum_{i=1}^{N} \ln \left( \sum_{k=1}^{K} \phi_k N(x_i; \mu_k, \Sigma_k) \right) \\
\text{s.t. } \begin{cases} 
\sum_{k=1}^{K} \phi_k = 1 \\
0 \leq \phi_k \leq 1.
\end{cases}
\]  

(4)

However, with the growth of spectral bands, performing density estimation and anomaly target detection in the original high-dimensional spectrum space is difficult for GMM since any spectrum may be a rare instance which resides in low-density areas. Hence, GMM may result in a suboptimal solution. To alleviate the curse of dimensionality, two-step approaches, in which representation learning is performed to conduct dimensionality reduction and anomaly target detection is performed on the extracted representations subsequently, are separately learned. However, these two-step approaches always fail to preserve the inherent information customized for anomaly target detection, since the representation learning model lacks the guidance from the subsequent detection procedure. Thus, these approaches may be stuck in suboptimal solutions easily. Thus, E2E-LIADE is proposed for digging out characterization customized for HAD and learning a low-dimensional representation that preserves the inherent information from the original high-dimensional spectral space, which enforces the following constraints:

\[
\begin{align*}
\min_{\theta_{en}, \theta_{de}, \Theta} & \quad \frac{1}{L} \sum_{i=1}^{L} \| x_i - \hat{x}_i \|^2 + \lambda_1 \frac{L}{L} \sum_{i=1}^{L} \sum_{j} \omega_{ij} \| c_i - c_j \|^2 \\
- \lambda_2 & \sum_{i=1}^{L} \ln(f(c_i; \Theta)) \\
\text{s.t. } & \quad c_i = \text{En}(x_i; \theta_{en}) \\
& \quad \hat{x}_i = \text{De}(c_i; \theta_{de}) \\
& \quad \sum_{j} \phi_k = 1 \\
& \quad \Theta = \{\phi, \mu, \Sigma\}
\end{align*}
\]  

(5)

where the first two components of the objective function denote a representation learning model, which ensures preserving the inherent information from the original high-dimensional spectral space, and the third component of the objective function represents a density estimation procedure. Furthermore, a joint optimization strategy and an estimation network, which lead to solving the optimization problem in (5) in an E2E manner, are proposed to dig out the characterization customized for HAD and approximate the optimal detection performance. More details of the proposed E2E-LIADE will be described in the following sections.

### A. Local Invariant Autoencoder

In hyperspectral data, two observations motivate us to propose a fundamental module LIA to explore the critical information for density estimation and anomaly target detection: 1) anomalies are spectrally distinct from their neighbors and 2) anomalies always occur with low probability and in small proportions, and therefore, contribute less to training.

According to observation 1), the local structural information of hyperspectral data can be adopted to characterize anomalies. Inspired by manifold learning, the intrinsic dimension of background spectra is considered to be much lower than the number of bands in hyperspectral data. Thus, an LIA, which regularizes the latent space to preserve the local geometry of hyperspectral data, is proposed to capture the intrinsic low-dimensional manifold embedded in the original high-dimensional space. In the learned manifold, the LIA will shorten the distance between spectra with similar spectral signatures and push away those with different spectral signatures.

The architecture of the proposed LIA is illustrated in Fig. 1. Both the encoder \( \text{En}(\cdot) \) and the decoder \( \text{De}(\cdot) \) are multilayer
fully connected neural networks with the tanh activation function. The spectra vectors \( \mathbf{x}_i \in \mathbb{R}^{B \times 1} \) \((i = 1, 2, \ldots, L)\) sampled from the original hyperspectral dataset \( \mathbf{X} \) are adopted as training data for LIA to learn the low-dimensional representation of the data. The encoder \( \text{En}(\cdot) \) aims to map the input sample \( \mathbf{x}_i \) as a low-dimensional representation \( \mathbf{c}_i \in \mathbb{R}^H \) with the parameter \( \theta_{en} \)

\[
\mathbf{c}_i = \text{En}(\mathbf{x}_i; \theta_{en}) = \tanh(\mathbf{W}_{en}\mathbf{x}_i + \mathbf{b}_{en})
\]

\[
\theta_{en} = (\mathbf{W}_{en}, \mathbf{b}_{en})
\]  

(6)

and the decoder \( \text{De}(\cdot) \) is used to reconstruct \( \mathbf{x}_i \) with \( \mathbf{c}_i \)

\[
\hat{\mathbf{x}}_i = \text{De}(\mathbf{c}_i; \theta_{de}) = \tanh(\mathbf{W}_{de}\mathbf{c}_i + \mathbf{b}_{de})
\]

\[
\theta_{de} = (\mathbf{W}_{de}, \mathbf{b}_{de})
\]  

(7)

where \( \theta_{de} \) is the parameter of \( \text{De}(\cdot) \).

Since the local structural information between spectra cannot be revealed by the original autoencoder constrained by the universal fidelity loss function, under the manifold learning assumption, the local invariance is introduced to constrain the representation learning. In detail, the local invariance assumes that if two spectra \( \mathbf{x}_i \) and \( \mathbf{x}_j \) are close to each other in the original high-dimensional space \( \mathbb{R}^{B \times 1} \), then the latent representations \( \mathbf{c}_i \) and \( \mathbf{c}_j \) should also be close to each other in the learned manifold. To achieve this objective, a locally invariant constraint is adopted as an extra regularizer for the loss function of the original AE

\[
\min_{\theta_{en}, \theta_{de}} \frac{1}{L} \sum_{i=1}^{L} \left\| \mathbf{x}_i - \hat{\mathbf{x}}_i \right\|^2 + \frac{\lambda_1}{L} \sum_{i=1}^{L} \sum_{j \neq i} \omega_{ij} \left\| \mathbf{c}_i - \mathbf{c}_j \right\|^2
\]  

(8)

where \( \lambda_1 \) is the weight coefficient of the local invariant constraint, and \( \omega_{ij} \) is the \((i, j)\)th entry of an affinity matrix \( \Omega = (\omega_{ij}) \in \mathbb{R}^{L \times L} \) and represents the feature similarity between input samples \( \mathbf{x}_i \) and \( \mathbf{x}_j \). Note that the corresponding low-dimensional representations \( \mathbf{c}_i \) and \( \mathbf{c}_j \) of \( \mathbf{x}_i \) and \( \mathbf{x}_j \) should have the same feature similarity as \( \mathbf{x}_i \) and \( \mathbf{x}_j \). The affinity matrix \( \Omega \), which plays an important role in analyzing the complex intrinsic relationships of the hyperspectral data, can be easily captured from the original data distribution. Besides, only a few nearest neighbors of a certain sample are taken into consideration to perform the local invariant constraint; hence, \( \Omega \) is a sparse matrix. More details about the affinity matrix \( \Omega \) will be discussed in Section II-D. With the constraint of local invariance, the learned representations in deep latent space can maintain the same geometrical structure as the original hyperspectral data. Furthermore, the loss function can be reformulated as the following matrix form:

\[
\min_{\theta_{en}, \theta_{de}} \frac{1}{L} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F + \frac{\lambda_1}{L} \text{tr}(\mathbf{C}\mathbf{M}\mathbf{C}^T)
\]  

(9)

where \( \text{tr}(\cdot) \) refers to the trace of a matrix, and \( \mathbf{M} \) is a Laplacian matrix constructed by

\[
\mathbf{M} = \mathbf{D}_1 + \mathbf{D}_2 - 2\mathbf{\Omega}
\]  

(10)

where \( \mathbf{D}_1 \) and \( \mathbf{D}_2 \) are diagonal degree matrices, whose \( i \)th diagonal entry corresponds to the summation of all the similarities related to \( \mathbf{x}_i \)

\[
\mathbf{D}_1 = \text{diag} \left( \sum_{j \neq i} \omega_{ij} \right)_{L \times L}
\]

(11)

\[
\mathbf{D}_2 = \text{diag} \left( \sum_{i \neq j} \omega_{ij} \right)_{L \times L}
\]  

(12)

Thus, the second term in (8) can be rewritten as

\[
\sum_{i=1}^{L} \sum_{j \neq i} \omega_{ij} \left\| \mathbf{c}_i - \mathbf{c}_j \right\|^2 = \sum_{i=1}^{L} \mathbf{c}_i^T \sum_{j} \omega_{ij} \mathbf{c}_j + \sum_{j=1}^{L} \mathbf{c}_j^T \sum_{i} \omega_{ij} \mathbf{c}_i
\]

\[
- 2 \sum_{i=1}^{L} \mathbf{c}_i^T \omega_{ij} \mathbf{c}_j
\]

\[
= \text{tr}(\mathbf{C}\mathbf{D}_1\mathbf{C}^T) + \text{tr}(\mathbf{C}\mathbf{D}_2\mathbf{C}^T)
\]

\[
- 2\text{tr}(\mathbf{C}\mathbf{\Omega}\mathbf{C}^T)
\]

\[
= \text{tr}(\mathbf{C}\mathbf{M}\mathbf{C}^T).
\]  

(13)

The LIA can be optimized as the loss function by the stochastic gradient algorithm. As shown in Fig. 1, with the locally invariant constraint, the geometrical structure of \( \mathbf{X} \) can be preserved in the local invariant low-dimensional representation \( \mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L] \).

However, only considering one of the observations may lead to a suboptimal solution. Based on observation 2), the reconstructed version of anomalies may exhibit high reconstruction errors. Taking full advantage of the complementary information provided by the local invariant low-dimensional representations and the spectral reconstruction errors can significantly enlarge the discrimination between anomalies and the background. In this article, a novel E2E multidistance measure, derived from the reconstruction error \( \mathbf{r}_i \), is proposed as

\[
\mathbf{r}_i = \begin{pmatrix} \text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i) \\ \text{OPD}(\mathbf{x}_i, \hat{\mathbf{x}}_i) \end{pmatrix} \in \mathbb{R}^{2 \times 1}
\]  

(14)

where \( \text{MSE} \) and \( \text{OPD} \) are defined as

\[
\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \frac{1}{B} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T (\mathbf{x}_i - \hat{\mathbf{x}}_i) \in \mathbb{R}
\]

(15)

\[
\text{OPD}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = (\mathbf{x}_i^T P_\perp \mathbf{x}_i + \hat{\mathbf{x}}_i^T P_\perp \hat{\mathbf{x}}_i) \in \mathbb{R}
\]  

(16)

with \( P_\perp = \mathbf{I} - \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \) for \( \mathbf{x} = \mathbf{x}_i, \hat{\mathbf{x}}_i \) and \( \mathbf{I} \) being a \( B \times B \) identity matrix. Thus, with the extracted local invariant low-dimensional representation \( \mathbf{c}_i \) and reconstruction error feature \( \mathbf{r}_i \), the ALDR \( \mathbf{z}_i \in \mathbb{R}^{H+2} \) of the original spectra vector \( \mathbf{x}_i \) can be generated by concatenating \( \mathbf{c}_i \) and \( \mathbf{r}_i \)

\[
\mathbf{z}_i = \begin{pmatrix} \mathbf{c}_i \\ \mathbf{r}_i \end{pmatrix} \in \mathbb{R}^{(H+2) \times 1}.
\]  

(17)

Finally, the LIA feeds the ALDR \( \mathbf{z}_i \) to the subsequent estimation network.
B. Estimation Network

Given the ALDR \( z_i \) and the number of mixture components \( K \), the mixture-component membership \( \hat{y}_i \in \mathbb{R}^{K \times 1} \) for the spectral vector \( x_i \) is predicted by the estimation network \( \text{Est}(\cdot; \theta_{es}) \)

\[
\hat{y}_i = \text{Est}(z_i; \theta_{es}) = \text{softmax}(\text{NN}(z_i; \theta_{es}))
\]

where \( \theta_{es} \) refers to the parameters of the neural network \( \text{NN} \). Given a batch of \( l \) spectra, \( \{x_1, \ldots, x_l\} \), and their mixture-component membership, the parameters of the density estimation procedure in (5) can be further estimated as

\[
\begin{align*}
\hat{\phi}_k &= \sum_{i=1}^{l} \frac{\hat{y}_{ik}}{l} \\
\hat{\mu}_k &= \frac{\sum_{i=1}^{l} \hat{y}_{ik} z_i}{\sum_{i=1}^{l} \hat{y}_{ik}} \\
\hat{\Sigma}_k &= \frac{\sum_{i=1}^{l} \hat{y}_{ik} (z_i - \hat{\mu}_k)(z_i - \hat{\mu}_k)^T}{\sum_{i=1}^{l} \hat{y}_{ik}}
\end{align*}
\]

where \( k = 1, 2, \ldots, K \), and \( \hat{\phi}_k, \hat{\mu}_k, \) and \( \hat{\Sigma}_k \) are mixture probability, mixture mean vector, and the mixture covariance matrix for the component \( k \), respectively. The probability density (3) of the spectral vector \( x_i \) can be derived with the estimated parameters

\[
p(z_i) = \sum_{k=1}^{K} \frac{\hat{\phi}_k}{\sqrt{2\pi |\hat{\Sigma}_k|}} e^{-\frac{1}{2} (z_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (z_i - \hat{\mu}_k)}
\]

where \( |\cdot| \) refers to the determinant of a matrix. Furthermore, the corresponding negative log likelihood of (22) can be formulated as

\[
E(z_i) = - \log \left( \sum_{k=1}^{K} \frac{\hat{\phi}_k}{\sqrt{2\pi |\hat{\Sigma}_k|}} e^{-\frac{1}{2} (z_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (z_i - \hat{\mu}_k)} \right).
\]

Based on the assumption that anomalous spectra lie in a low-probability density area of an augmented low-dimensional space, it is straightforward to identify spectra with low-probability density as anomalies. Thus, as shown in Fig. 1, an initial energy map \( E \in \mathbb{R}^{M \times N} \), where each value refers to the anomalous degree of the corresponding spectrum, can be derived by calculating the negative log likelihood \( E(z_i) \).

C. Joint Optimization Strategy

As shown in Fig. 1, to avoid local optima, during the training procedure, the LIA and the estimation network are jointly optimized in an E2E manner. With the GD-based joint optimization procedure, the LIA can effectively preserve the ALDR with the guidance of the trainable objective customized for HAD, and the estimation network can make meaningful density estimations with the well-learned ALDR from the LIA. Thus, the LIA and the estimation network can mutually promote each other’s performance. The proposed E2E-LIADE can be mathematically formulated as the following optimization problem:

\[
\min_{\theta_{en}, \theta_{de}, \theta_{es}} \frac{1}{L} \sum_{i=1}^{L} \| x_i - \hat{x}_i \|^2 + \lambda_1 \frac{1}{L} \text{tr}(CMC^T) + \lambda_2 \sum_{i=1}^{L} E(z_i) + \lambda_3 g(\hat{\Sigma})
\]

Here, \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are the hyperparameters, which are set as 0.1, 3, and 0.0001 in practice, respectively. The objective function of the optimization problem includes four components. The first component refers to the reconstruction error caused by the LIA. The second component is a regularizer that preserves the local structural information of the hyperspectral data. The third \( E(z_i) \) characterizes the probability density of the observed input samples in the augmented low-dimensional space. The optimal combination of LIA and estimation network, which maximize the likelihood of the observed input samples, can be established by minimizing (24). Furthermore, when the diagonal entries of covariance matrices degenerate to 0, trivial solutions are produced. Thus, by introducing the fourth component, small values on the diagonal entries can be penalized and trivial solutions can be avoided

\[
g(\hat{\Sigma}) = \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{1}{\hat{\Sigma}_{jj}}
\]

where \( d \) is the size of the covariance matrix \( \hat{\Sigma} \).

D. Graph-Based Local Invariant Regularizer Design

A graph \( G \) with \( n \) vertices is constructed, in which each vertex represents a spectrum. The affinity matrix \( \Omega = (\omega_{ij}) \in \mathbb{R}^{L \times L} \) is adopted as a symmetric weight matrix, where \( \omega_{ij} \) is the weight of the edge joining vertices \( i \) and \( j \). Thus, a weighted undirected graph \( G = (V, U, \Omega) \) is constructed with the weight matrix \( \Omega \). Here, \( V = \{v_i\}_{i=1}^{n} \) is the vertex set in which each node \( v_i \) refers to a spectrum, and \( U = \{u_{ij}\} \) is the edge set where each edge \( u_{ij} \) associates nodes \( v_i \) and \( v_j \) with a weight \( \omega_{ij} \). The values of \( \omega_{ij} \) are defined as

\[
\omega_{ij} = \begin{cases} 
1, & \text{if } x_i \in N_{3}(x_j) \text{ or } x_j \in N_{3}(x_i) \\
0, & \text{otherwise}
\end{cases}
\]

where \( \delta \) is a local invariant parameter, and \( N_{3}(x_i) = \{ x \mid |x - x_i| < \delta, x \in X \} \), that is, \( N_{3}(x_i) \) contains all the spectra whose distances to \( x_i \) are less than \( \delta \).

E. Postprocessing With Background Suppression

On the basis of the observation that anomalies always exhibit distinguished intensity and tiny size properties in the spatial domain, the initial energy map \( E \in \mathbb{R}^{M \times N} \) obtained by the proposed estimation network is refined with a postprocessing procedure to suppress background objects with large areas as illustrated in Fig. 3. The initial energy map \( E \) is processed as

\[
S(i, j) = \begin{cases} 
E(i, j), & \text{if } B(i, j) = 1 \\
E(i, j)^T, & \text{if } B(i, j) = 0
\end{cases}
\]
Fig. 3. Schematic of postprocessing with background suppression.

Fig. 4. Shape of nonlinear grayscale transformation $E(i, j)^\Gamma$ with different choices of $\Gamma$.

where $S(i, j)$, $E(i, j)$, and $B(i, j)$ refer to the pixels at the $i$th row and the $j$th column in $S \in \mathbb{R}^{M \times N}$, $E \in \mathbb{R}^{M \times N}$, and $B \in \mathbb{R}^{M \times N}$, respectively. $B$ is a Boolean map in which the connected components whose area is larger than $\beta$ are assigned 0 by the function $B W_{\text{area fitting}}(\cdot, \beta)$

$$B = BW_{\text{area fitting}}(T, \beta)$$

where $T \in \mathbb{R}^{M \times N}$ is the binary version of the initial energy map $E$ obtained by

$$T(i, j) = \begin{cases} 1, & E(i, j) > \xi \\ 0, & E(i, j) \leq \xi. \end{cases}$$

Here, $T(i, j)$ refers to the pixel at the $i$th row and the $j$th column in $T$. The parameter $\xi$ is chosen with the adaptive thresholding algorithm Otsu [36]. There is no influence on the detection performance when $\beta$ is too large. Alternately, both background and anomalies can be suppressed with a very small $\beta$ and thus weaken the detection performance. With a proper $\beta$, the background objects with large areas and noise in the initial energy map $E$ can be suppressed by the nonlinear grayscale transformation $E(i, j)^\Gamma$ in which $\Gamma$ is a positive parameter for shape adjustment as shown in Fig. 4, while the anomaly targets can be preserved, such that the detection performance can be enhanced. In this article, the optimum values for $\Gamma$ and $\beta$ are selected as 5 and $(N/100)$, respectively, in which $N$ is the number of spectra in the HSI. Furthermore, the guided filter [37] is adopted to refine $S$ and obtain the background suppressed detection map $D$. As illustrated in Fig. 3, the background can be suppressed properly.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets

Five real-world hyperspectral datasets, which are described as follows, are utilized to verify the accuracy and effectiveness of our proposed E2E-LIADE.

1) San Diego Dataset: With two corresponding images, the San Diego dataset, acquired by the airborne visible/infrared imaging spectrometer (AVIRIS) sensor covering the area of San Diego, is widely used in many research. The San Diego-1 data have 189 spectral bands in total covering the wavelengths ranging from 370 to 2510 nm and 100 $\times$ 100 spatial pixels. In the scene, hangars, parking aprons, and exposed soil constitute the main background, and three airplanes with 134 spatial pixels are regarded as anomalies. The San Diego-2 has 193 spectral bands varying from 370 to 2510 nm, along with 100 $\times$ 100 spatial pixels. In this scene, the main background materials are beach and water. The man-made objects with 202 spatial pixels in the water are selected as the anomalies. The pseudocolor images and ground truth are shown in Fig. 5(a) and (b).

2) Cat Island Dataset: With 100 $\times$ 100 spatial pixels and 193 spectral bands, the Cat Island dataset was also acquired by the AVIRIS sensor covering the Cat Island area. Noisy bands in the original images have been removed. It has 100 $\times$ 100 spatial pixels and 193 spectral bands. The main background are the sea and island, while a ship with 19 spatial pixels in this scene is regarded as anomaly. The pseudocolor image and ground truth are presented in Fig. 5(c).

3) Gainesville Dataset: The dataset was acquired by the AVIRIS sensor over the area of Gainesville city. The image scene covers an area of 100 $\times$ 100 pixels, with 191 spectral bands varying from 430 to 860 nm, with a 3.5-m spatial resolution. There are some ships considered as anomalies, which, in total, consists of 52 pixels.

4) HYDICE Dataset: Acquired by the hyperspectral digital imagery collection experiment (HYDICE) sensor over an urban area, CA, USA, the HYDICE dataset has 100 $\times$ 80 spatial pixels with 162 spectral channels. The spectral resolution is 10 nm, and the spatial resolution is 1 m. It consists of an urban area covering vegetation, buildings, roads, and vehicles. The anomalies are vehicles of different sizes composed of 19 pixels in the urban scene. The 2-D visualization of the HYDICE dataset and its corresponding ground truth is depicted in Fig. 5(e).

B. Experimental Setup

1) Comparison Methods: Six HAD methods are adopted for comparison, that is: 1) RX; 2) LRX; 3) CRD; 4) LRASR; 5) LSMAD; and 6) AED. In the experiments, the parameters
of the LRX, CRD, LRASR, LSMAD, and AED methods are selected optimally for each dataset according to [15], [17], [18], and [21], respectively.

2) Performance Evaluation Metrics: In the field of signal detection, the receiver operating characteristic (ROC) [38] curve and the area under the ROC curve (AUC) [39] serve as valid criterions to evaluate the superiority and effectiveness of the proposed method qualitatively and quantitatively. As a widely used evaluation criterion in HAD, the ROC curve can be generated by three parameters, that is: 1) true positive rate (TPR); 2) false positive rate (FPR); and 3) threshold (τ). TPR and FPR are the probability of detection and probability of the false alarm, respectively. The ROC curve of (TPR, FPR) describes the varying relationship between TPR and FPR. The AUC is used to evaluate the detection performance quantitatively. The larger the AUC of (TPR, FPR), the better the detection accuracy. The smaller the AUC of (FPR, τ), the lower the false alarm rate. Furthermore, the separability between anomaly and background is evaluated via Box-Whisker Plots. The Box-Whisker plot displays the variation in samples of a statistical population, which is non-parametric without any assumption of the underlying statistical distribution. The degree of dispersion and skewness in the data can be described by the spacings between the different parts of the box. More details about the Box-Whisker plot can be found in [40].

3) Implementation Details: The proposed E2E-LIADE is composed of the LIA and estimation network, and each part is composed of a two-layer fully connected network. The number of hidden nodes in the deep latent space is set to 8. We train E2E-LIADE with the SGD optimizer [41] in an E2E fashion, setting the learning rate to $10^{-4}$ and the batch size to 5. Intuitively, it is reasonable to estimate the distribution of complex HSIs with the larger parameter $K$; however, it takes too much time to convergence. In E2E-LIADE, we empirically set the number of Gaussian mixture components $K$ to 8 considering the tradeoff between detection performance and training time. We terminate the learning process in 1000 epochs. The hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\delta$ are empirically set as 3, 0.1, 0.0001, and 0.01, respectively. We implement our method by TensorFlow-GPU 1.15 on one NVIDIA 2080Ti GPU with 11-GB memory.

C. Parameter Sensitivity Analysis

This section is an concrete analysis of the sensitivity of the parameters on the performance of the proposed method. $H$ is set in the feature extractor to represent the number of hidden nodes, $K$ is the number of Gaussian components, and $\lambda_1$, $\lambda_2$, and $\lambda_3$ in (24) are the hyperparameters. The optimal parameter settings for the proposed method are selected based on experience. In this experiment, when the analysis of a certain parameter is performing, the other parameters are fixed as the corresponding optimal values of $H = 8$, $K = 5$, $\delta = 0.01$, $\lambda_1 = 3$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.0001$. As the subfigures are
shown in Fig. 6, each horizontal axis is the parameter, and the vertical axis shows the AUC scores of (TPR, FPR).

1) **Number of Hidden Nodes** \( H \): For the number of hidden nodes in the feature extractor, that is, LIA net, the value is ranging from 6 to 10 for all the datasets, to analyze the sensitivity of the proposed method. As shown in Fig. 6(a), for a smaller value of \( H \), the LIA net is unable to obtain a discriminative representation for the original spectral data, while a larger value of \( H \) may make the LIA net unable to suppress the redundant information and the noise in the original HSI, degrading the final detection performance. With a proper \( H = 8 \), the LIA net can extract a good reduced representation for the input spectral and promote the detection performance.

2) **Number of Gaussian Components** \( K \): For the number of Gaussian components, the value is varying from 1 to 10. The estimation net can effectively characterize the distribution of hyperspectral data and separate anomalies from the background with a proper parameter \( K \). As illustrated in Fig. 6(b), the AUCs tend to increase as the value of parameter \( K \) increases and then become stable until \( K \) is over 5. In addition, it is intuitively plausible that the distribution of a hyperspectral data with complex ground objects can be well estimated with a larger parameter \( K \).

3) **Local Invariant Parameter** \( \delta \): For the local invariant parameter, the value is varying from \( 10^{-6} \) to 1. \( N_{\delta}(x_i) = \{ x \mid |x - x_i| < \delta, x \in X \} \) contains all the spectra whose distances to \( x_i \) are less than \( \delta \). For a larger \( \delta \), some anomalies may be mixed with the background instances in the embedded low-dimensional manifold, while a smaller value of \( \delta \) cannot capture the instinct structure of the hyperspectral dataset. As illustrated in Fig. 6(c), with the proper value of \( \delta = 10^{-2} \), the model can fully explore the critical information for anomaly target detection and lead to satisfactory detection performance.

4) **Hyperparameters** \( \lambda_1, \lambda_2, \text{ and } \lambda_3 \): As shown in (24), there are four components and three hyperparameters (i.e., \( \lambda_1, \lambda_2, \text{ and } \lambda_3 \)) in the objective function. The relationship among the three hyperparameters can be characterized as 1: \( \lambda_1 : \lambda_2 : \lambda_3 \). A larger value of \( \lambda_1 \) may lead to ineffective learning of low-dimensional representation, while a smaller value of \( \lambda_1 \) may lose the local invariant information. A proper value of \( \lambda_2 \) can ensure the estimation net to converge to an ideal state so that the distribution of the data can be well estimated. Meanwhile, with the proper value of \( \lambda_3 \), the singularity effect can be countered and the model can be stable during the training procedure. As illustrated in Fig. 6(d)–(f), \( \lambda_1 = 3, \lambda_2 = 0.1, \text{ and } \lambda_3 = 0.0001 \) are selected as the optimal hyperparameter settings.

**D. Ablation Study**

In order to explore the effectiveness of each proposed module on detection performance, we assess the detection quality of different modules, including the proposed LIA, proposed estimation network, and joint optimization procedure. The first ablation study pointed out that the original AE determines anomalies by calculating reconstruction errors between the original spectra and the decoded ones. GMM-EM estimates the probability of anomalies in the original high-dimensional space by the traditional EM algorithm. In this case, samples that are distributed in low-probability density regions are more
likely to be anomalies. AE+GMM-EM (two step) is a separate two-step approach. At the first step, the original AE extracts the latent features of the hyperspectral dataset. At the second step, an alternative two stage algorithm, that is, the EM algorithm, is adopted to take the latent features from the AE and train GMM. AE-EstNet (joint optimize) and LIA-EstNet (joint optimize) are both E2E models. The former jointly optimizes an original AE and an estimation net to perform feature extraction and anomaly target detection. The latter jointly optimizes the proposed LIA and an estimation net to exploit the embedding manifold and perform anomaly target detection.

Table I summarizes the results of ablation studies. For the original AE, only considering the reconstruction error is insufficient. For the GMM-EM, due to the curse of dimensionality, performing rational density estimation is difficult. For the AE+GMM-EM (two step), the critical information specific to anomaly target detection may be lost during the feature extraction step. Compared with AE+GMM-EM (two step), AE-EstNet (joint optimize) and LIA-EstNet (joint optimize) are both E2E models. The former jointly optimizes an original AE and an estimation net to perform feature extraction and anomaly target detection. The latter jointly optimizes the proposed LIA and an estimation net to exploit the embedding manifold and perform anomaly target detection.

Table IQUANTITATIVE RESULTS FOR ABLATION STUDY

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Average AUC of (TPR, FPR)</th>
<th>Average AUC of (FPR, (\tau))</th>
<th>Computing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original AE</td>
<td>0.8862</td>
<td>0.0469</td>
<td>11.54</td>
</tr>
<tr>
<td>Original GMM-EM</td>
<td>0.9608</td>
<td>0.0858</td>
<td>55.14</td>
</tr>
<tr>
<td>AE+GMM-EM (two-step)</td>
<td>0.9616</td>
<td>0.0481</td>
<td>56.35</td>
</tr>
<tr>
<td>AE-EstNet (joint-optimize)</td>
<td>0.9624</td>
<td>0.0831</td>
<td>59.98</td>
</tr>
<tr>
<td>LIA-EstNet (joint-optimize)</td>
<td>0.9970</td>
<td>0.0129</td>
<td>64.60</td>
</tr>
</tbody>
</table>

E. Detection Performance

The detection maps are shown in Fig. 5. Tables II and III, and Fig. 7 indicate the AUC of (TPR, FPR), the AUC of (FPR, \(\tau\)), and the ROC curves, respectively.

For San Diego-1, the proposed method can well detect aircrafts in the airstrip with the highest intensity and a relatively low false alarm. The LRX, LRASR, and AED methods can identify the position of the anomaly, with the shapes of anomalies failing to be preserved. Meanwhile, for the RX, CRD, and LSMAD methods, edge details of some anomalies are mixed with the background due to high false alarm rate. For San Diego-2, the anomalies in the sea can be detected by the proposed method and AED with a relatively low false alarm rate. The RX, LRASR, and LSMAD methods also identify all the anomalies but exhibit a high false alarm rate. The LRX can detect all the anomalies but with low intensity. At the same time, for the AED method, several pixel-level anomalies are blurry. The RX, LRASR, and LSMAD methods also identify all the anomalies but exhibit a high false alarm rate. Meanwhile, the CRD method almost lost all the anomalies visually. Tables II and III indicate that the proposed method yields better detection results than other methods and

![Fig. 7. ROC curves for five hyperspectral dataset. (a) SanDiego-1. (b) SanDiego-2. (c) Cat Island. (d) Gainesville. (e) HYDICE.](image-url)
achieves the highest AUC of (TPR, FPR) for San Diego-1 and San Diego-2.

For the Cat Island dataset, the RX, LRASR, LSMAD, and AED can detect the anomaly while they yield a high false alarm rate at the lower left part of the scene. The CRD method suppresses the background interference but cannot preserve the edge details of the target. Both LRX and the proposed method can retain the targets in the sea with a high intensity and a low background interference.

For the Gainesville dataset, as shown in Fig. 5(d), we notice that both the proposed method and AED can well highlight anomaly objects with different sizes and shapes, and suppress the complex backgrounds. However, the LRX and CRD miss anomalies to different degrees, while the RX, LRASR, and LSMAD are more likely to be confused by the background.

Fig. 5(e) displays the detection results on HYDICE dataset. CRD methods almost lost all the anomalies visually. As to the detection results of RX and LSMAD methods, most anomalies are mixed with the background visually. The LRASR and AED yield better results but the suppression of background interference is not effective. Both LSMAD and the proposed method yield a superior detection result with better background suppression, while the proposed method exhibits the best detection performance in terms of quantitative criterion and visual analysis. The AUC values on the HYDICE dataset are shown in Table II, and our proposed method presents a relatively low AUC of (FPR, τ) of 0.0110 as shown in Table III.

To further evaluate the ability of the aforementioned methods to distinguish anomalies from the background, the Box-Whisker plots corresponding to different datasets are plotted in Fig. 8. The detection values of each pixel are calculated, and boxes are plotted to enclose the main parts of the pixels, excluding the biggest 10% and the smallest 10%. The anomaly and background columns for each detector are plotted as red boxes and blue boxes, respectively. The lines at the top and bottom of each column are the extremums, which are normalized to 0–1, and the line in the middle of the box is the mean of the pixels. The position of the boxes reflects the tendency and compactness of the distribution of the pixels. In other words, the separability of anomalies and background is indicated by the position. For the San Diego dataset shown in Fig. 8(a) and (b), the proposed method successively increases the separation between the anomaly box and the background box. The anomaly box is overlapped with the background box for RX and LRX. Meanwhile, the boxes are overlapped for CRD in San Diego-2. Similarly, for the San Diego-1, there are hardly any gap between anomalies and background boxes for LSMAD and AED. In terms of the Cat Island dataset, Fig. 8(c) indicates that all the HAD methods have an obvious separation between the anomaly box and the background box. Moreover, the proposed method can more effectively suppress the background interference in comparison with the others. For the Gainesville dataset, Fig. 8(d) illustrates that only the proposed method can ensure the gap between the anomaly box and the background box. For the HYDICE dataset, Fig. 8(e) indicates that LRX and CRD cannot ensure the gap between the anomaly box and the background box. However, the gaps between the two boxes for LSMAD, AED, and the proposed E2E-LIADE are notable, and the performance of the proposed method is more satisfactory.

Besides, in order to evaluate the computational burden, as illustrated in Table IV, we further tabulate the actual runtime of different methods. Although the proposed method is not the most efficient method, it is much faster than LRX and CRD. All the experiments have been carried out on an NVIDIA 2080Ti GPU with 8-GB memory.

### IV. Conclusion

In this article, a novel and unified E2E-LIADE framework for unsupervised HAD is presented. The key point of our method is that an input spectral vector directly corresponds to an anomaly density estimation through unsupervised learning, which aims at achieving an optimal solution while satisfying the assumption on manifold. Accordingly, the superiority of E2E-LIADE is resulted from several aspects. The LIA is developed to capture the intrinsic low-dimensional manifold embedded in the original high-dimensional space, and
ALDR is generated by concatenating the local invariant low-dimensional representation solved by the graph regularizer and the proposed novel multidistance measure, which reveals the reconstruction error. Then, an estimation network estimates the sample distribution on ALDR, and an energy map of hyperspectral data is generated. A joint optimization strategy is performed on several real datasets demonstrate that the proposed E2E-LIADE exhibits a satisfactory performance in terms of quantitative evaluation and visual comparison.

REFERENCES