Hyperspectral Anomaly Detection Based on Spatial–Spectral Cross-Guided Mask Autoencoder

Qing Guo, Yi Cen, Lifu Zhang, *Senior Member, IEEE*, Yan Zhang, and Yixiang Huang

***Abstract*—Autoencoders (AEs) have gained widespread application in the field of hyperspectral anomaly detection, largely due to their notable effectiveness in efficiently reconstructing backgrounds within hyperspectral images (HSIs). However, the absence of prior knowledge and constraints imposed by spectral information capacity hinder the accuracy of anomaly detection by allowing AEs to reconstruct both anomalous targets and backgrounds simultaneously. To address this limitation, a spatial–spectral cross-guided masked autoencoder (SSCMAE) has been proposed. The guided mask is generated based on the spectral difference between the anomaly and the background. This mask effectively suppresses the reconstruction of anomalous targets while enhancing the accuracy of background reconstruction. Moreover, a dual-branch structure operates, encompassing spatial and spectral dimensions, effectively capturing the inherent three-dimensional characteristics present in HSIs. Ingeniously designed cross-connection layers within the architecture enhance the spatial and spectral branches' capability of extracting internal spatial and spectral features of images. In order to capture a more comprehensive range of background features, a lightweight three-dimensional convolutional autoencoder is introduced. This addresses the issue of local feature loss during background reconstruction and overcomes the limitations that visual transformers face when learning local image structures. The proposed method has been systematically compared against several advanced methods on six real-world datasets. The results explicitly demonstrate the efficacy and superior performance of the presented SSCMAE approach.**

***Index Terms*—Anomaly detection, autoencoder (AE), cross connect, guided mask, hyperspectral image (HSI).**

I. Introduction

HYPERSPECTRAL images (HSIs), which consist of multiple contiguous bands and provide detailed spectral information, offer distinct advantages for hyperspectral target detection [1], [2]. Hyperspectral anomaly detection is a method used for target detection without prior knowledge [3]. An anomaly refers to an object that significantly deviates from the background in spatial or spectral terms. It is characterized by low probability and small size [4], [5]. Presently, hyperspectral anomaly detection is of great importance in a wide range of applications, including vegetation and agricultural monitoring, atmospheric research, geological exploration, marine studies, and military reconnaissance [6], [7], [8].

According to the basic methods employed, there are three main categories of hyperspectral anomaly detection methods: statistical feature-based, feature expression-based, and deep learning-based [9]. Statistical feature-based methods assume that the target or background follows specific mathematical statistical distributions. Among these methods, Reed–Xiaoli (RX) [10] stands out as the current benchmark approach to hyperspectral anomaly detection. The RX algorithm operates under the assumption that the background model conforms to a multivariate Gaussian distribution. The algorithm evaluates the anomaly status of each pixel by comparing the disparity between the hyperspectral data of the pixel in question and that of its surrounding pixels. To reduce the impact of anomalous pixels and enhance detection accuracy, various advanced variants have been developed. This has led to the creation of a set of advanced variants, such as localized RX [11] and clustering kernel RX [12], among others. Nevertheless, the intricate distribution of real HIS poses a challenge for simple models to effectively characterize the background distribution, thereby constraining the overall performance of anomaly detection.

To address the limitations of basic statistical distribution assumptions, representation-based methods have been introduced for hyperspectral anomaly detection. In these methods, a dictionary is used to reconstruct pixels within a specific model, and the residuals are then used to indicate the level of anomaly [9]. Typical algorithms include the collaborative representation-based detector (CRD) [13], the anomaly detection method based on low-rank and sparse representation [14] method, and the low-rank and sparse matrix decomposition-based Mahalanobis distance method [15]. However, these methods overlook the global structural information of HSI. Furthermore, creating a thorough background dictionary without interference from anomalous pixels is difficult without prior information. Moreover, when data dimensions are very high, the computational cost of expression-based methods becomes a significant limitation.

In contrast to the previously mentioned approaches, deep learning possesses the capability of extracting latent features from HSI, thus offering a comprehensive representation of the

Manuscript received 23 January 2024; revised 19 March 2024; accepted 17 April 2024. Date of publication 26 April 2024; date of current version 17 May 2024. This work was supported by the National Key Research and Development Project of China under Grant 2022YFF0904400. *(Corresponding author: Yi Cen.)*

Qing Guo, Yan Zhang, and Yixiang Huang are with the National Engineering Laboratory for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: guoqing22@mails.ucas.ac.cn; rs_zhangyan@163.com; huangyixiang20@mails.ucas.ac.cn).

Yi Cen and Lifu Zhang are with the National Engineering Laboratory for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: cenyi@radi.ac.cn; zhanglf@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3393995

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

intricacies in both spatial and spectral dimensions, as well as multidimensional features [16], [17], [18]. This capability enhances the distinction between background and anomalies. Unsupervised deep learning methods have gained significant attention in anomaly detection due to the lack of spectral information for both background and anomalous targets. An autoencoder (AE) is a classic unsupervised deep learning model that excels at revealing intrinsic data structures and patterns by capturing abstract and intricate feature representations from HSI [19], [20]. Its successful application in anomaly detection has produced outstanding results.

For example, He et al. [21] proposed a module initially based on clustering to detect pseudobackground and anomalous samples. This was followed by an integration of convolutional and transformer operations to extract both local and global discriminative features. Wang et al. [22] proposed the residual self-attention module to extract important features, reducing the ability of the subsequent network to reconstruct anomalies. Furthermore, it is assumed that the background in the original space possesses low-rank properties, which leads to the development of a low-rank loss function designed to suppress the reconstruction of anomalies. Wang et al. [23] introduced a new hyperspectral anomaly detection method called the dynamic negative sampling-based AE. This approach an adaptive-adjusted loss function to suppress the reconstruction error of original pixels while amplifying the error of negative samples. The incorporation of skip connections ensures that features from both shallow and deep layers are utilized in the reconstruction process.

It is not difficult to find that although AE effectively improves anomaly detection performance, it primarily focuses on the inherent spectral information of HSI, overlooking their spatial characteristics. This is evident in the oversight of crucial semantic details, such as intricate textures, and the failure to account for the correlation between pixel spatial features and HSI. In addition, most methods inevitably reconstruct the anomalies simultaneously, especially in situations where the background closely resembles the visual characteristics of the anomalies. Consequently, inhibiting the manifestation of the anomalies presents a significant challenge.

While AEs effectively improve anomaly detection performance, HSIs are three-dimensional, and anomalous objects typically manifest as small areas and distinct spectral features. Therefore, spatial features, in addition to spectral behavior, naturally become another valuable source of information. For instance, Zhao et al. [19] combined deep belief networks with spatial filtering to extract spectral and spatial features. Xie et al. [24] utilized structural tensor focusing on edges and corners, guiding filters to obtain initial detection maps. Tu et al. [25] employed graph Laplacian anomaly and differential fusion methods for global and local anomaly detection based on spectral adjacency matrices and spatial adjacency matrices. Recent studies have shown that utilizing both spatial and spectral information simultaneously is more effective than relying solely on spectral information. Acquiring spatial information is crucial for capturing semantic details, such as complex textures, and understanding the correlation between pixel spatial features and HSI. Most anomaly detection methods based on AE primarily focus on the intrinsic spectral information of HSI while overlooking its spatial feature anomalies. This approach leads to the simultaneous reconstruction of anomalies, especially in situations where background and anomaly visual characteristics closely resemble each other. Hence, inhibiting the reconstruction of anomalies presents a significant challenge.

To address the aforementioned issues, a method for hyperspectral anomaly detection called spectral–spatial cross-guided mask autoencoder (SSCMAE) has been proposed. In this method, the network reconstructs the background, and anomalies are identified as reconstruction errors. Specifically, first, considering the capability of masked image modeling to integrate global contextual information into local features, it captures a significant amount of latent image features as well as the spectral differences between anomalies and background. A guided mask has been meticulously designed to suppress the reconstruction of anomalous targets while simultaneously improving the accuracy of background reconstruction. SSCMAE adopts a dual-branch mask AE structure in spatial and spectral dimensions, effectively utilizing the inherent three-dimensional information in HSI to reconstruct backgrounds. The spatial branch uses guided and random masks to reconstruct the pixels that have been masked. The spectral branch focuses on reconstructing spectral channels using a random mask. Moreover, the carefully designed cross-connected layers enhance the spatial and spectral branches' ability to represent spatial and spectral features within the image. To further explore the capture of local background features and address the challenge of losing these features when queries are embedded in the attention layers of the visual transformer (ViT), a lightweight three-dimensional convolutional autoencoder (3DCAE) is used to learn local features. The primary contributions of this article can be outlined in the following three ways.

1) A new approach for identifying anomalies in HSI is presented, utilizing a guided mask AE. This method involves the use of a guided mask to protect the target, allowing the attention mechanism in the reconstructed network to focus solely on the background information.
2) A spectral–spatial cross-guided mask structure is used in the SSCMAE model to fully exploit the local features of the unmasked 3-D HSI and obtain global background information. The cross connection effectively compensates for the lack of detailed information in both spatial and spectral dimensions, thereby improving the accuracy of background reconstruction.
3) A hybrid network combining ViT and 3DCAE is proposed to extract local spatial features of HSI. This addresses the issue of missing local features in background reconstruction and overcomes the limitations of ViT in learning local structures.

The rest of this article is organized as follows. Section II provides an overview of related work. Section III delves into the details of the proposed method. Section IV showcases and analyzes the experimental results, and finally, Section V concludes the article.

II. Related Work

In this section, we concisely present two pivotal studies to our network: AEs and ViT.

A. Autoencoders

AE is a neural network consisting of an input layer (encoder), a latent layer, and a reconstruction layer (decoder) [18], [26]. It operates in an unsupervised manner, learning features from HIS by minimizing the reconstruction error at the decoding layer. The encoding layer captures deep features from the input data. The input layer is transformed into the hidden layer $\mathbf{Z}$ through the application of the weight matrix $\mathbf{W}$ and the summation bias $\mathbf{b}$

$$\mathbf{Z} = \sigma(\mathbf{W}\mathbf{I} + \mathbf{b}). \tag{1}$$

The decoding layer uses the deep features from the hidden layer $\mathbf{Z}$ to reconstruct the image $\mathbf{I}'$

$$\mathbf{I}' = \sigma(\mathbf{W}'\mathbf{Z} + \mathbf{b}') \tag{2}$$

where σ is the activation function, $\mathbf{W}'$ is the decoding layer weight matrix, and $\mathbf{b}'$ is the decoding layer bias.

B. Vision Transformer

The transformer architecture, originally introduced in the domain of natural language processing, has outperformed previous intricate recursive and convolutional neural network (CNN) models and has emerged as a seminal model in the field [16]. Dosovitski et al. [27] expanded the transformer model into the field of computer vision by introducing the ViT. Expanding on the original transformer encoder, ViT incorporates a CNN with a transformer model that encompasses global self-attention. This amalgamation allows the model to capture comprehensive contextual information in images, addressing the limitations of CNNs in modeling long-range dependencies and enhancing feature representation [17]. ViT has demonstrated success in various visual domains, including image classification [28], object detection [29], and semantic segmentation [30]. Over the past few years, there has been an increasing focus on the application of transformers in the field of hyperspectral anomaly detection. Wang et al. [31] elected typical background pixels for training a transformer-based AE, aiming to achieve the reconstruction of background pixels. Xiao et al. [32] utilized a spatial–spectral dual-window mask transformer to consolidate background information from a global perspective across the entire image. This was done to reduce irregularities, enable thorough feature extraction, and minimize abnormal reconstructions by using neighboring pixels within the dual-window framework.

III. Proposed Approach

In this section, a concise overview of the proposed research method is initially presented, followed by an in-depth discussion of the specific implementation aspects related to its individual components, culminating in a description outlining the general structure of the method.

A. Overview

Our proposed method comprises four primary elements: a masking strategy, background reconstruction facilitated by a spatial–spectral cross-masking AE, a module for local spatial feature extraction using a 3-D convolutional autoencoding network, and an anomalous target detection module. The architectural overview of SSCMAE is shown in Fig. 1. Initially, two strategic approaches, random masking and guided masking, are utilized on spatial and spectral masks, resulting in distinct mask maps. The unmasked maps are then input into the network to reconstruct the background. Considering that the transformer may lose local information during training, a spectral and spatial cross-linking mechanism is designed to address the deficiency of detailed information in the spatial and spectral dimensions. Subsequently, a 3-D convolutional autoencoding network is used to extract local information in both spatial and spectral dimensions to reconstruct the background. Finally, anomaly resulting maps are generated by applying the Mahalanobis distance.

B. Mask Strategy

HIS is distinguished by its multiband nature, large volume, and high redundancy. In the context of HSI, anomalous targets are irregularly embedded within the background, occurring at frequencies lower than those of background pixels. The strong learning capability of AEs enables them to effectively reconstruct anomalous targets, thereby influencing detection accuracy. Employing a masking strategy to reconstruct masked regions from adjacent unmasked areas facilitates better capture of local features and structures in the image. This process improves the quality of background reconstruction and enhances anomaly detection performance. However, current methods often use random masking for basic data augmentation, without taking into account the differences between anomalies and background. However, existing methods often use random masking without taking into account the distinctions between anomalies and backgrounds.

To address this limitation, the masking strategy is divided into two parts: random mask generation and guided mask generation. Masking is performed in two dimensions: spatial and spectral dimensions, respectively. As shown in Fig. 2. This dual approach aims to delve deeper into the correlation of HSI, extract more comprehensive and effective potential features from the background, and ultimately improve the quality of background reconstruction.

1) Guided Mask: Anomaly targets exhibit distinct characteristics: the background pixels show a strong resemblance to the average vector, while the anomalous pixels show a low resemblance to the average vector. Zhang et al. [33] proposed a framework for anomaly target detection. This framework integrates outlier removal through an iterative strategy. Expanding on Zhang's concept, the spectral angle between the mean vector and individual pixels in HSI serves as a criterion for identifying and removing suspected anomalous targets. The spectral angle between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as the cosine of the angle between them, which facilitates the comparison of spectral

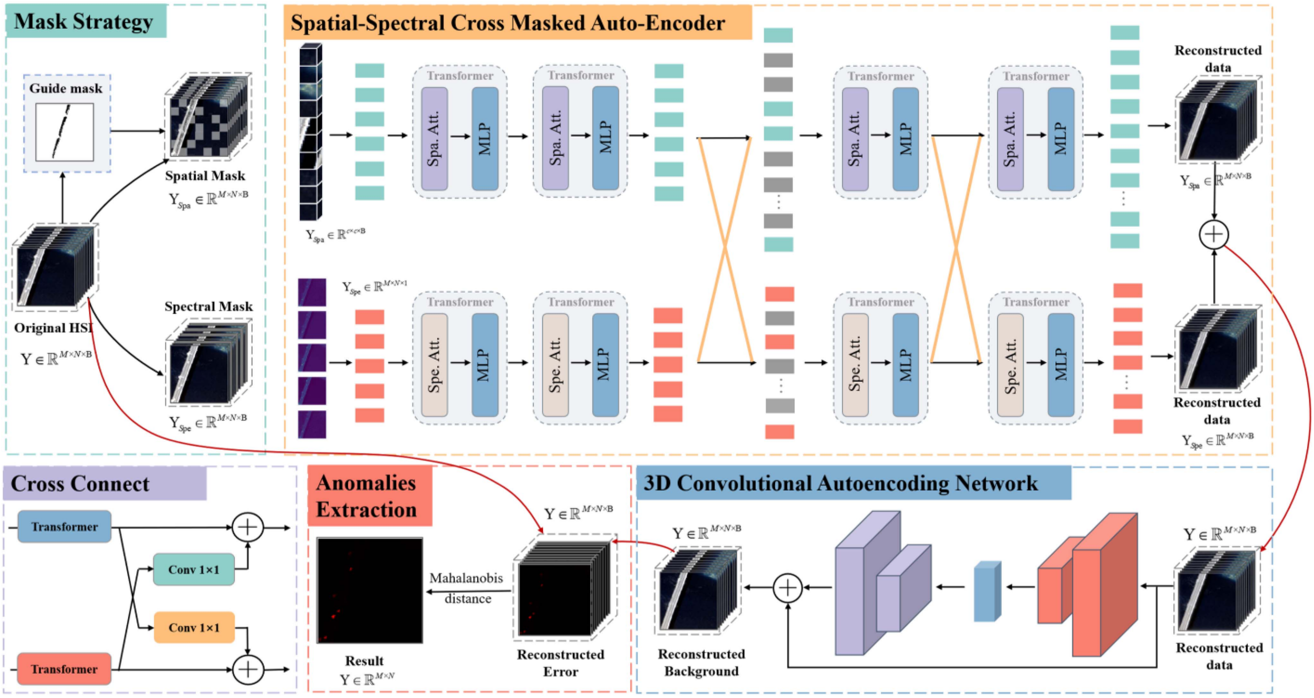

Mask Strategy
Guide mask
Spatial Mask
$Y_{Spa} \in \mathbb{R}^{M\times N\times B}$
Original HSI
$Y \in \mathbb{R}^{M\times N\times B}$
Spectral Mask
$Y_{Spe} \in \mathbb{R}^{M\times N\times B}$
Spatial-Spectral Cross Masked Auto-Encoder
$Y_{Spa} \in \mathbb{R}^{c\times c\times B}$
$Y_{Spe} \in \mathbb{R}^{M\times N\times 1}$
Transformer
Spa. Att.
MLP
Spe. Att.
Reconstructed data
$Y_{Spa} \in \mathbb{R}^{M\times N\times B}$
Reconstructed data
$Y_{Spe} \in \mathbb{R}^{M\times N\times B}$
Cross Connect
Transformer
Conv 1×1
Conv 1×1
Transformer
Anomalies Extraction
Mahalanobis distance
Result
$Y \in \mathbb{R}^{M\times N}$
$Y \in \mathbb{R}^{M\times N\times B}$
Reconstructed Error
3D Convolutional Autoencoding Network
$Y \in \mathbb{R}^{M\times N\times B}$
Reconstructed Background
$Y \in \mathbb{R}^{M\times N\times B}$
Reconstructed data

Fig. 1. Overall pipeline of the proposed SSCMAE method.

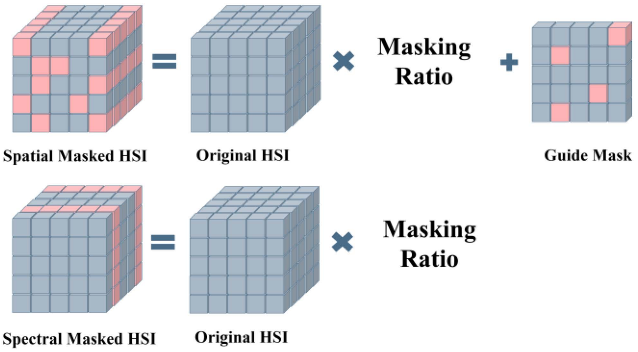


Fig. 2. Masking strategy involves spatial and spectral masking.

signatures

$$s(\boldsymbol{a}, \boldsymbol{b}) = \cos^{-1}\left(\frac{<\boldsymbol{a}, \boldsymbol{b}>}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|}\right) \tag{3}$$

where <***a***,***b***> denotes the dot product of vectors ***a*** and ***b***. $\|a\|$ and $\|b\|$ represent the magnitudes (Euclidean norms) of vectors ***a*** and ***b***, respectively. This formula calculates the angle (in radians) between two vectors in a high-dimensional space. The cosine of this angle is used as a measure to quantify spectral similarity.

To reduce the sensitivity of the Z-Score algorithm to outliers, a method for outlier removal based on the Z-Score is proposed. The Z-score value reflects how much a data point deviates from the mean of the entire dataset. For each data point, the mean and standard deviation of the dataset are calculated. Subsequently, each data point is transformed into its corresponding Z-Score value Z_i

$$Z_i = \frac{x_i - \mu}{\sigma} > \text{threshold} \tag{4}$$

where x_i represents the data value, μ represents the mean of the dataset, and σ represents the standard deviation of the dataset. Anomalies, which are significantly distant from the mean, are identified through larger spectral angles. These outlier values, which are identified by their deviation, are flagged for removal. The *threshold* is set at three times the standard deviation, which is easy to use and widely accepted to identify outliers.

2) Spatial Masking: Initially, guided masking is used to obscure certain pixels in the spatial dimension. However, due to the significant redundancy of pixel information in the image, the model can use simple interpolation to approximate the masked pixels. This makes it challenging for the model to understand the higher level meaning of the image. Consequently, random masks are applied to obscure certain pixels, facilitating the extraction of meaningful global information. The HSI is divided into overlapping $(H \times W)/(M \times M)$ patch blocks. For each of these patch blocks, both a guided mask and a random mask are applied $x_p = \mathbb{R}^{r\times M\times M\times C}$. Here, *r* represents the mask ratio, wherein the masked region is assigned a value of 0, while the remaining pixels are configured to 1.

3) Spectral Masking: Owing to the high correlation between neighboring bands in HSI, which results in increased redundancy, extracting the appropriate spectral information becomes challenging. Instead of solely focusing on reconstructing image blocks, the emphasis is shifted toward extracting useful spectral information from a subset of bands. To accomplish this, band information is masked. The HIS $\mathrm{I} \in \mathbb{R}^{H\times W\times C}$ is then

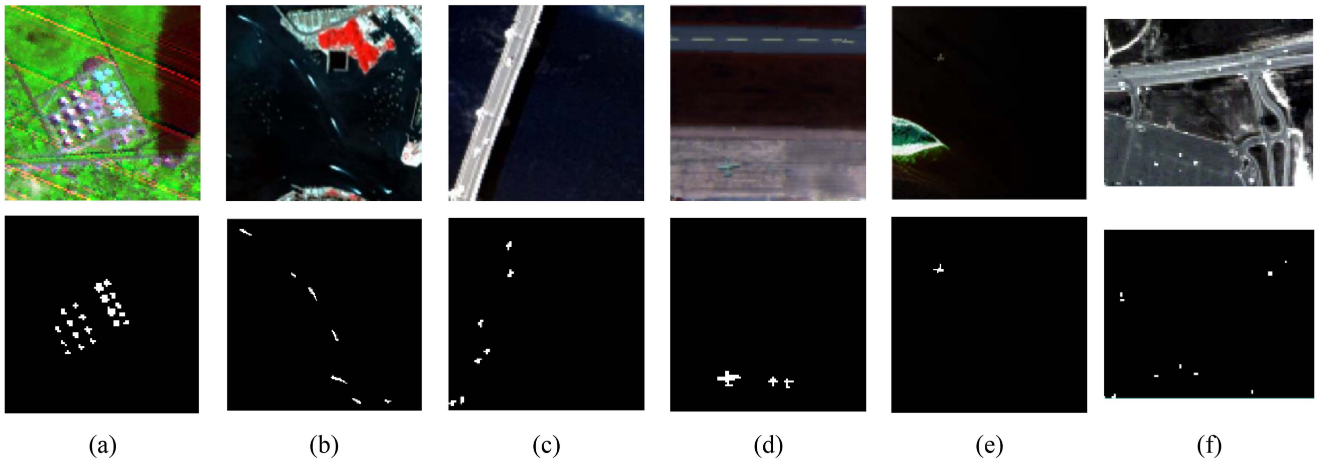


Fig. 3. False-color images and ground-truth images of the experimental datasets. (a) Texas coast. (b) Belcher bay. (c) Pavia. (d) Gulfport. (e) Cat Island. (f) HYDICE.

reshaped into a series of equally-sized, nonoverlapping 2-D images I_{Spectral}, totaling $(H \times W \times C)/(M \times M)$, where M is the patch size. Each channel block can be treated as a 2-D grayscale image. The encoder processes a series of these patches $x_p \in \mathbb{R}^{r \times C \times M \times M}$, r representing the mask ratio.

The masking ratio is an important factor that influences the results of reconstruction. Therefore, comprehensive experiments were conducted in Section IV-C to determine the optimal masking rate.

C. Spatial–Spectral Cross-Attention Transformer

1) Encoder–Decoder: The encoder follows a standard ViT model but exclusively accepts unmasked image blocks as input. These visible image blocks undergo a linear mapping to include positional embeddings before being processed by the transformer block for feature extraction. The transformer module consists of a multihead self-attention module (MHSA) and a multilayer perceptron (MLP). The same MLP architecture as suggested in ViT is used [27]. The MHSA can be seen as a function that uses linear projection to transform inputs into different spaces in order to obtain Query, Key, and Value. It then calculates attention scores through scaled dot product, and the final output is derived by a weighted sum and linear transformation

$$\hat{\mathbf{X}} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{X} \tag{5}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k})\mathbf{V} \tag{6}$$

where $\mathbf{X}$ and $\hat{\mathbf{X}}$ are the input and output feature maps. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are reshaped tensors derived from the input. $\sqrt{d_k}$ learnable scaling parameter is utilized to regulate the softmax. To execute multihead attention, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are divided into h heads along the feature channel dimension. This enables the efficient and parallel learning of separate attention maps.

In spatial attention, the input features are augmented augmentation through a 1 × 1 convolution. Following this, the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are derived. Spectral attention $x_p = \mathbb{R}^{r \times M \times M \times C}$ is defined in a similar manner to spatial attention, with the distinction that the spectral dimensions are transformed after linear projection to a dimension denoted as $\boldsymbol{z}_0 = [x_p^1 E, x_p^2 E \cdots, x_p^{(H \times W)/(M \times M)} E]$. where E represents the number of feature dimensions for each band after linear projection, and it signifies the similarity between the spectral bands being focused on.

Corresponding to the encoder structure, the decoder also includes a transformer module. The input to the decoder consists of three elements: encoded visible patches, mask markers, and cross connects. The mask marker indicates a learnable shared vector, signifying that the image block at that location needs to be reconstructed. Cross connections facilitate the exchange of spectral and spatial feature information. Following this, the resulting data are sent to the transformer module to undergo the process of acquiring a comprehensive understanding of the intricate feature mapping. Once the decoding process is finalized, the dimensions of the feature mapping are aligned with those of the initial image.

2) Cross Connect: The self-attention mechanism in ViT tends to prioritize global information, which may limit its ability to effectively capture local features and result in a loss of detailed information in reconstruction tasks [34]. In contrast, the designed cross-connected layers efficiently share information between tasks. Configured as an assembly of fundamental units (illustrated by the dashed rectangle in Fig. 1), each unit (n) receives an input mapping and transmits the output mapping to the subsequent unit ($n + 1$). Each unit consists of a core layer tailored for a specific task (depicted by blue and pink blocks) and an auxiliary convolutional layer that connects two CNNs (illustrated by green and orange blocks). The connected convolutional layer has a kernel size that is the same as the number of channels in the output mapping of the alternative stream. These interconnections serve the purpose of transferring information rather than extracting features. Effectively compensating for the lack of detailed information in both spatial and spectral

TABLE I
PARAMETER SETTINGS OF THE PROPOSED 3D-CAE WHEN APPLIED TO DATA BY THE TEXAS COAST IMAGE

	Kernel size	Strides
Conv1+ Bn1	2×3×3×32	1×1×1×1
Conv2+ Bn2	2×3×3×64	1×1×1×1
Deconv1+ Bn3	2×3×3×32	1×1×1×1
Deconv2+ Bn4	2×3×3×32	1×1×1×1

dimensions, these connections play a crucial role in incorporating additional discriminative information to distinguish diverse features. Designating the input maps for the *n*th unit as $\boldsymbol{x}_n^A$ and $\boldsymbol{x}_n^A$, and *t* denoting the transformations learned by the original convolutional layers as f_n^A and f_n^B, respectively. Assuming the cross-connection layers learn transformations g_n^A and g_n^B, then $\boldsymbol{x}_{n+1}^A$ and $\boldsymbol{x}_{n+1}^B$ are computed as follows:

$$\begin{aligned}\boldsymbol{x}_{n+1}^A &= f_n^A(\boldsymbol{x}_n^A) + g_n^A(f_n^B(\boldsymbol{x}_n^B))\\ \boldsymbol{x}_{n+1}^B &= f_n^B(\boldsymbol{x}_n^B) + g_n^B(f_n^A(\boldsymbol{x}_n^A)).\end{aligned} \tag{7}$$

The transformations $f_n^B(\boldsymbol{x}_n^B)$ and $f_n^A(\boldsymbol{x}_n^A)$ are derived from the encoder. Subsequently, $\boldsymbol{x}_{n+1}^A$ and $\boldsymbol{x}_{n+1}^B$ are used in the decoder.

D. 3-D Convolutional Autoencoding Network

ViTs, relying on MHSA, excel in establishing long-distance models for comprehensive global sensory field coverage [31]. However, they may lack the capability of capturing local details similar to CNNs, potentially resulting in the loss of localized information such as edges and textures. To address this, a lightweight 3DCAE is proposed based on a 3-D convolutional neural network (3DCNN). The 3DCNN treats the HSI cube as a unified entity, allowing for accurate and efficient extraction of features from the combined deep null-spectrum of HSIs. This process directly yields 3-D feature cubes, which represent a significant improvement over previous models that were based on 1-D, 2-D, or 2-D + 1-D. The approach not only utilizes spatial and spectral information but also takes into account spatial–spectral correlation. Importantly, it achieves this with fewer parameters and layers, thereby enhancing efficiency in feature extraction.

3DCAE is an advanced neural network model based on the conventional AE. In our model, a relatively shallow neural network is developed specifically to exploit the characteristics of anomalous data. This network consists of two 3-D convolutional layers serving as the encoder and two 3-D deconvolutional layers serving as the decoder, without any pooling or fully connected layers included. A 3×3×2 convolutional kernel with a stride of 1 is chosen to maximize the extraction of spatial and spectral features, as well as enhance noise suppression capabilities. In addition, 3-D batch normalization (BN) is employed as a regularization technique to normalize the features generated by each 3-D convolutional layer. This ensures consistency in feature weight ranges and helps mitigate the risk of overfitting. The architecture of the 3DCAE used in this experiment is described in Table I.

```
Algorithm 1: SSCMAE.
 Input: hyperspectral image $\mathbf{H}$
 Initialization: window size $c$, masking ratio $r$,
   multiattention mechanism heads $h$, and feature dimension
   size $d$
   Calculate the guided mask $\mathbf{M}$ according and create
      spatial mask and spectral mask according to Mask
      Strategy;
   Reconstruct the HSI background $\hat{\mathrm{H}}$
   Extracting localized information about the
      reconstructed background
   Extract the anomaly targets
 Output: Extract the anomaly targets $\mathbf{\Delta H}$
```

E. Extraction of Anomaly Targets

The mask integrates the reconstructed background in both spectral and spatial dimensions, linearly combining the two-dimensional reconstructed backgrounds

$$D = \omega_1 D_{\text{Spectral}} + \omega_2 D_{\text{Spatial}} \tag{8}$$

where ω_1 and ω_2 serve as balancing parameters. Without any prior knowledge, it is challenging to ascertain whether anomalies in the dataset manifest at the pixel level or within regions containing structural information. Drawing inspiration from [35], setting $\omega_1 = \omega_2 = 0.5$ indicates that spectral and spatial features contribute equally to anomaly detection in the absence of a priori information.

To train the proposed SSCMAE network, a conventional backpropagation algorithm is employed. The chosen loss function for training is the mean squared error, defined as follows:

$$\mathcal{L} = \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 \tag{9}$$

where $\mathbf{x}$ is the original image, and $\tilde{\mathbf{x}}$ is the reconstructed image.

Upon the completion of training in the SSCMAE network, the reconstructed background of the HIS is obtained. Mahalanobis distance serves as an effective metric for assessing the similarity between sample groups. By utilizing the covariance matrix to represent distance, the Mahalanobis distance can adapt to the correlations between variables. This property enables the Mahalanobis distance to amplify the influence of small changes in variables and effectively identify anomalies [36]. The detection result can be expressed as follows:

$$D(\mathbf{x}) = (\mathbf{x} - \mu)\Gamma^{-1}(\mathbf{x} - \mu)^{\mathrm{T}} \tag{10}$$

where $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n]$ is an *n*-dimensional HSI pixel vector; μ and Γ are the mean and the covariance matrix of the input background data.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed methodology undergoes testing on six authentic datasets shown in Fig. 3 and is systematically compared against cutting edge approaches. The experimental results are subjected to a comprehensive examination using both qualitative and quantitative methodologies. Furthermore, a comprehensive analysis of the method is conducted to investigate the impact of parameters, such as input patch size and shielding rate. In addition, a comprehensive ablation study is conducted to demonstrate the effectiveness of the proposed method.

A. Experimental Datasets

1) Texas Coast Image: This dataset captures and depicts the Texas coast in the United States using the airborne visible/infrared imaging spectrometer (AVIRIS) [37]. The image is 100 × 100 pixels with 207 spectral bands and a spatial resolution of 17.2 m. The primary anomaly is a building, which contrasts against the predominantly vegetated background.

2) Belcher Bay Image: This dataset captures the hyperspectral imager aboard the GF-5 satellite, providing a detailed view of Belcher Bay in Hong Kong, China. The image is 150 × 150 pixels with 150 spectral bands, and it boasts a spatial resolution of 30 m. Notable anomalies within the image include ships and ship tracks, set against a background of water bodies, structures, and vegetation. The presence of finely fragmented rocks in the water adds a layer of complexity, which could potentially result in false alarms during analysis.

3) PaviaC Image: This dataset captures the city center of Pavia in northern Italy, obtained through a Reflective Optical System Imaging Spectrometer (ROSIS-03) sensor [38]. The imagery is presented in a 100 × 100-pixel format, featuring 102 spectral bands with a spatial resolution of 1.3 m. Anomaly objects within this urban context are represented by cars on the bridge, set against the background of the bridge and river.

4) Gulfport Image: This dataset captures Gulfport in America, acquired by AVIRIS with a 3.4 m spatial resolution [39]. This 100 × 100-pixel image consists of 191 spectral bands. The anomalies in this HSI dataset are uniquely portrayed by three airplanes positioned at the bottom of the image, which are almost unrecognizable to the naked eye in the false-color image.

5) Cat Island Image: This dataset captures Cat Island in Japan, acquired by AVIRIS with a 17.2 m spatial resolution [40]. The image, sized at 150 × 150 pixels with 188 spectral bands, depicts a background of the sea and an island. Notably, a ship on the sea is identified as an anomaly, creating a captivating contrast within the maritime setting.

6) HYDICE Urban Image: This dataset depicts a suburban residential area in Michigan, USA, captured by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor with a 3 m spatial resolution [9]. The image, sized at 80 × 100 pixels with 175 spectral bands, depicts a background of vegetation, soil, water, and road. Ten man-made vehicles are identified as anomalies. Notably, to make detection more difficult, we retained the bands of the water absorption regions with low signal-to-noise ratio and poor quality (1–4, 76, 87, 101–111, 136–153, and 198–210).

B. Parameter Analysis

The impact of parameter configurations on anomaly detection performance is explored for the task of reconstructing the background in the mask. Two types of parameters were fine-tuned for the network structure (including the transformer feature dimension size and the number of multiattention mechanism heads) and the input data (comprising the input patch size, spatial masking ratio, and channel masking ratio). To scrutinize a specific hyperparameter, all other hyperparameters were maintained at a constant value. The default values were set as window size $c = 5$, masking ratio $r = 0.3$, multiattention mechanism heads $h = 4$, and feature dimension size $d = 512$. The model is optimized using AdamW with a learning rate set to 0.001 and a weight decay set to 1.

1) Token Dimension: The Token Dimension in the Transformer model refers to the dimensionality of hidden states at each position, which influences the model's expressiveness and comes with associated computational and memory costs. The optimal selection of the hidden layer dimension requires a thorough evaluation of both model performance and resource constraints.

Across all datasets, parameters c, r, and h were consistently configured to 5, 0.3, and 4, respectively. The parameter d was systematically varied within the range of 32 to 512. The impact of the variable d on the performance of the proposed method is illustrated in Fig. 4(a). The optimal feature size for Cat Island and HYDICE is 256, while all other datasets are 512. This shows that the appropriate feature dimension for each data can effectively extract the best semantic information.

2) Heads Number: The number of multiattention mechanism heads allows for simultaneous focus on various input aspects, thereby enhancing the model's ability to capture intricate details. While satisfactory performance can often be achieved with a modest number of variables, increasing the number of variables may lead to superior results when dealing with complex datasets.

Across all datasets, the parameters c, r, and d were consistently set to 5, 0.3, and 512, respectively. The parameter h was systematically varied within the range of 1 to 16. The influence of the variable h on the performance of the proposed method is depicted in Fig. 4(b). The optimal feature size for all datasets is 4. Elevating the number of heads, denoted as h, does not yield a substantial improvement in performance and, in certain instances, may even result in a decline in accuracy.

3) Patch Size: Adjusting the patch size has implications for the balance between global information and details. Enlarging the patch size enables the capture of more global information at the cost of fine details, while reducing the patch size aids in preserving finer details but may compromise the inclusion of comprehensive global features.

The window size was observed to influence the quality of image reconstruction. Across all datasets, the parameters h, r, and d were fixed at 4, 0.3, and 512, respectively. The parameter c was systematically adjusted within the range of 5–13, and its impact on the performance of the proposed method is depicted in Fig. 4(c). The Belcher Bay, Texas Coast, PaviaC, Gulfport, Cat Island, and HYDICE datasets require optimal patch Size of 5, 13, 13, 7, 5, and 9, respectively.

The proposed method involves reconstructing the masked portion using the unmasked portion. Therefore, the choice of patch size significantly impacts the context information of the mask. The six datasets can be categorized into two groups: the Texas Coast, PaviaC, and HYDICE. These datasets consist of

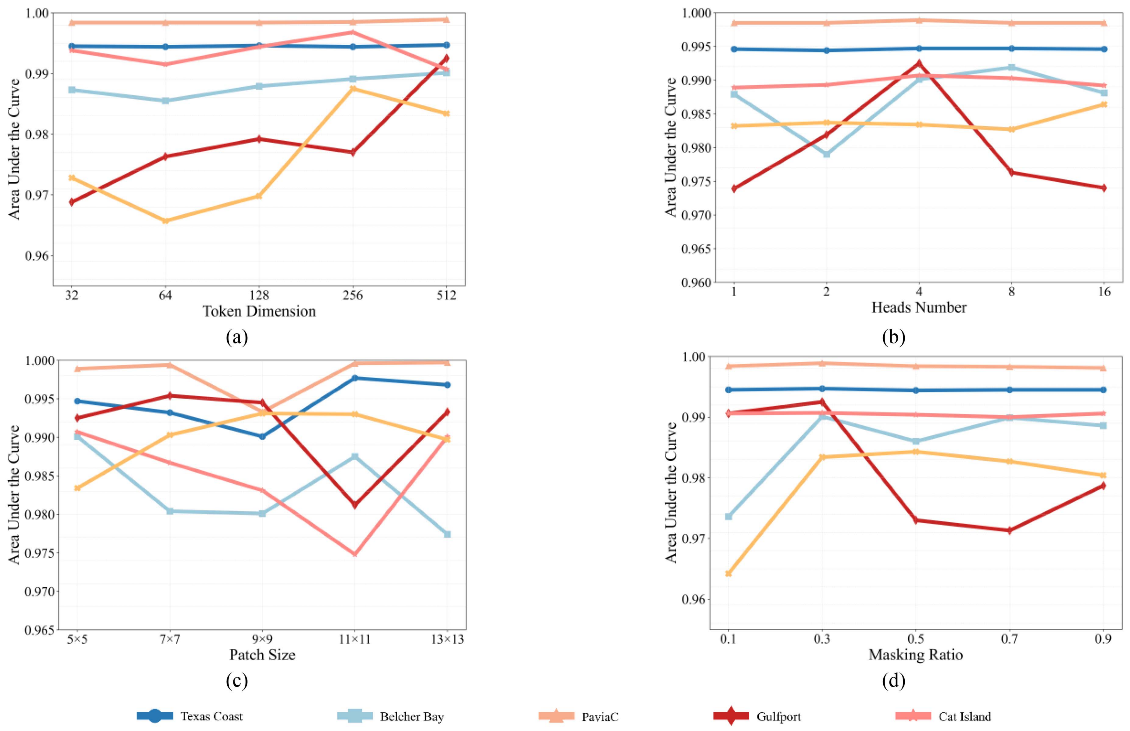


Fig. 4. Effect of different parameters on the performance of the proposed method. (a) Token dimension d. (b) Head number h. (c) Patch size c. (d) Spatial/channel masking ratio r.

small targets with high density. In this instance, the larger patches are adequate to cover the entire target without sacrificing important information. In the other three datasets, which comprise sparse, large targets, the smaller patches are useful for capturing localized features.

4) Spatial/Channel Masking Ratio: An appropriate masking ratio contributes to meaningful feature learning, robustness, and prevention of overfitting. However, an excessively high masking rate may restrict the model's ability to learn crucial features. Conversely, a too low masking rate might result in memorizing training data excessively, neglecting noise, and subsequently impacting the model's generalization on new data. In practice, choosing an appropriate masking rate involves finding a balance between model complexity, data diversity, and training convergence speed. This requires experimentation and tuning to determine the optimal value.

The window size was identified as a factor influencing the quality of image reconstruction. Across all datasets, parameters h, c, and d were fixed at 4, 5, and 512, respectively. The parameter r was systematically adjusted within the range of 0.1–0.9, and its impact on the performance of the proposed method is illustrated in Fig. 4(d). Notably, the proposed method exhibited relative stability across various values of r. Ultimately, except for the HYDICE dataset where r is set to 0.5, for other datasets, r is set to 0.3. This is attributed to the presence of numerous low signal-to-noise ratios and poor-quality bands in the HYDICE dataset, necessitating a higher masking ratio to effectively extract meaningful information. The results indicate that high ratios may pose challenges for the model in restoring the full range of information, potentially impacting detection performance. Therefore, it is essential to carefully choose the parameter r to ensure optimal image reconstruction and detection capabilities.

In conclusion, the backgrounds of the PaviaC, Texas Coast, and Cat Island datasets are simpler and cleaner, while the backgrounds of the Belcher Bay, Gulfport, and HYDICE datasets are more challenging to distinguish from the spectra of the anomalies. The anomalies consist of dozens of pixels and are larger in size, making the effects of the different hyperparameters more pronounced on them.

C. Experimental Setup

1) Evaluation Metrics: In the experiments, the detection performance of the proposed method was evaluated and compared with other methodologies using metrics, such as the three-dimensional receiver operating characteristic (3D-ROC) curve, the area under the curve (AUC) [41], and statistical separability maps [19]. The experiment utilizes 3D-ROC with two types of two-dimensional ROC curves: ROC (PD, PF) and ROC (τ, PF). ROC (PD, PF) describes the relationship between the false positive rate (PF) and the true positive rate (PD) at different thresholds τ. The PF–τ curve is used to evaluate the false alarm probability. The AUC value corresponds to a quantitative measure used to assess the accuracy of detection. In an ideal scenario, a detector would produce an ROC curve that is closely positioned to the upper-left corner, resulting in an AUC value approaching 1. Furthermore, separability maps are used to evaluate the discernibility between anomalies and the background. The boxes shown on the plot represent the distribution range

TABLE II
PARAMETER SETTINGS OF VARIOUS ALGORITHMS

Detector	Parameter	Texas Coast	Belcher Bay	PaviaC	Gulfport	Cat Island	HYDICE
LRX	(Win, Wout)	(19, 21)	(5, 7)	(7, 9)	(11, 25)	（5, 7）	(5, 3)
CRD	(Win, Wout)	(19, 21)	(5, 7)	(7, 9)	(17, 21)	(7, 9)	(5, 3)
GTVLRR	Beta	1000	0.01	1000	0.01	0.01	0.01
	Lambda	0.01	0.01	0.01	0.1	0.1	0.01
	Gamma	0.01	0.001	0.001	0.01	0.01	0.01
LEBSR	p-norm: 0.5						
NJCR	Lambda:1000, sigma: [0.25,0.5]						
GAED	learning rate: 0.1, penalty coefficient: 1, number of iterations: 300						
Auto-AD	learning rate: 0.01, number of iterations: 1001						

of detection values for both anomalies and the background. Consequently, the observed separation between these two boxes indicates the degree of distinguishability between anomalies and the background.

2) Comparison Methods: To evaluate the effectiveness of the proposed methodology, experimental comparisons are conducted with both conventional and deep learning-based approaches. The traditional methods include statistics-based methods (RX [10] and LRX [11]) and representation-based methods (CRD [13], GTVLRR [42], LEBSR [43], and NJCR [44]). The deep learning-based methods include autoencoder-based methods (GAED [45] and Auto-AD [20]). The experiment determines the uncertain parameters for cutting edge methods based on either the recommendations of the original authors or the AUC value. RX is omitted from the table since they do not require parameter settings. The parameters are set as given in Table II. Given the inherent uncertainty in detection outcomes, particularly in deep learning-based methodologies, the optimal result obtained from 20 consecutive runs for each method across the six experimental datasets is presented. Furthermore, all experiments are conducted on a machine equipped with an Intel Core i7-10700 central processing unit and 64 GB of random access memory. The learning and inference processes, using the network model, are performed within a Python 3.8 and PyTorch 2.0.1 environment, while other operations are conducted in the MATLAB R2018b environment.

D. Experimental Results

Fig. 5 illustrates the visual detection maps for all experimental datasets, Fig. 6 displays the corresponding 3D-ROC curves, and Fig. 7 displays separability maps obtained by different detectors on each dataset. In Table IV, the AUC values for the compared methods are calculated, and the second results are highlighted and underlined. Notably, the best results are presented in boldface.

Consistently, in Fig. 5, it is evident that the proposed SSCMAE achieves a commendable balance between anomaly detectability and background suppressibility as its corresponding detection maps closely align with the ground truth. Using the Pavia dataset as an illustrative example, the CRD struggles to discern the shape of the anomaly. In contrast, RX, LRX, GTVLRR, NJCR, and GAED demonstrate relatively effective performance in highlighting anomalous targets. However, these methods also tend to preserve certain background structural information, which can result in a higher frequency of false alarms. Compared to other algorithms, LEBSR and Auto-AD exhibit a robust ability to suppress background information. However, it also suppresses anomaly information, which fails to adequately reflect the shapes of the anomalies. In contrast, our proposed SSCMAE addresses these issues and accurately identifies anomalies and their shapes, consequently leading to a lower false alarm rate.

Table III presents the AUC values for the nine algorithms. It is evident that the proposed method outperforms six conventional and cutting-edge algorithms, consistently maintaining stable detection performance across all six datasets. This outcome highlights the significant advantage of SSCMAE over traditional and cutting-edge algorithms in effectively characterizing HSI.

Fig. 6 illustrates the 2-D and 3-D ROC curves based on nine methods across six datasets. As depicted in Fig. 6(a1)–(a6), it can be observed that the proposed SSMAE method consistently outperforms other methods as the false alarm rate $\mathrm{P_f}$ increases, particularly evident in the Gulfport image where SSMAE exhibits overwhelming superiority over other methods. It is noteworthy that although the proposed method visually outperforms other methods only within a limited range for the Belcher Bay and HYDICE datasets, utilizing the logarithmic scale of the x-axis demonstrates its superiority over a broader range. This indicates that our SSMAE demonstrates excellent overall detection performance compared to RX, LRX, CRD, GTVLRR, NJCR, LEBSR, Auto-AD, and GAED methods. As shown in Fig. 6(b1)–(b4), the two-dimensional ROC (PF, τ) curves are presented. Although the false alarm rate of SSMAE is higher than that of statistical methods (RX and LRX), its ability to suppress false alarms exceeds that of expression-based methods (such as CRD, GTVLRR, and NJCR) and is comparable to deep learning-based methods (such as GAED and Auto-AD) on most datasets. In addition, the three-dimensional ROC curves in Fig. 6(c1)–(c4) indicate that SSMAE exhibits superior performance across different datasets.

To further illustrate the effectiveness of SSCMAE in separating anomalous targets and suppressing background,

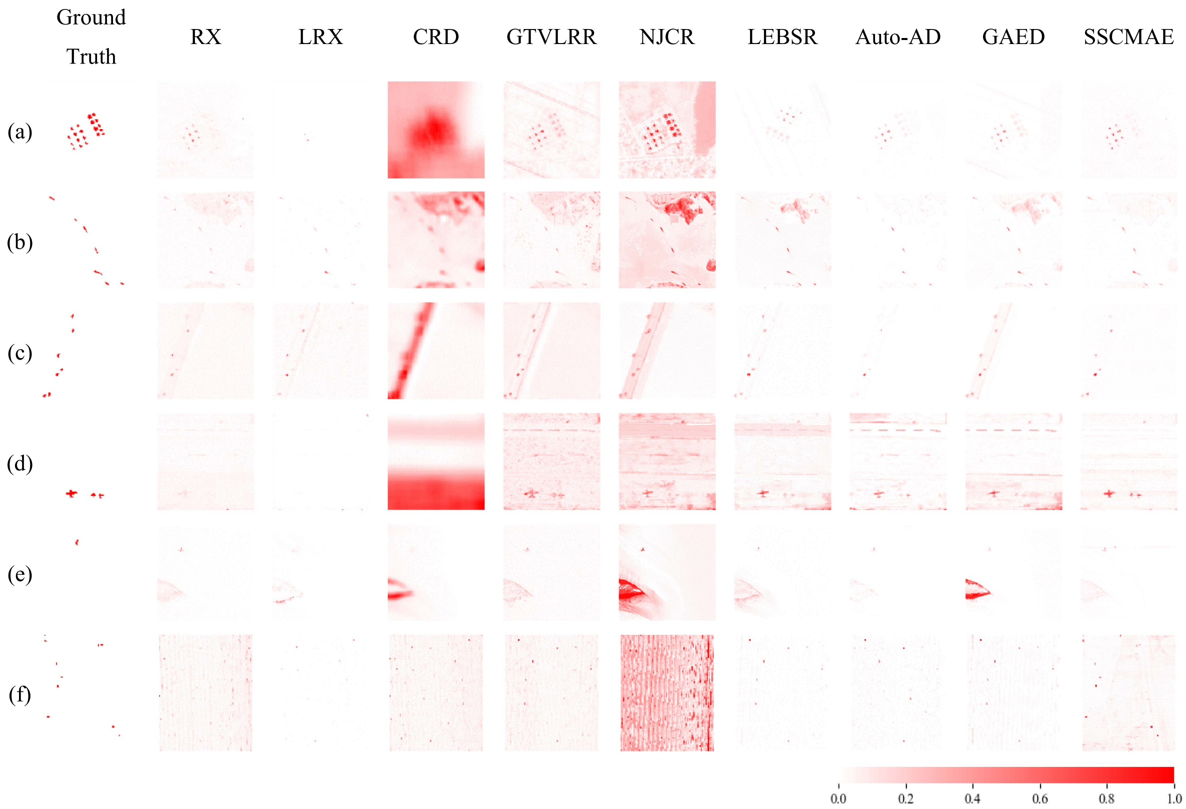


Fig. 5. Detection results of different methods on each dataset. (a) Texas coast. (b) Belcher bay. (c) Pavia. (d) Gulfport. (e) Cat Island. (f) HYDICE.

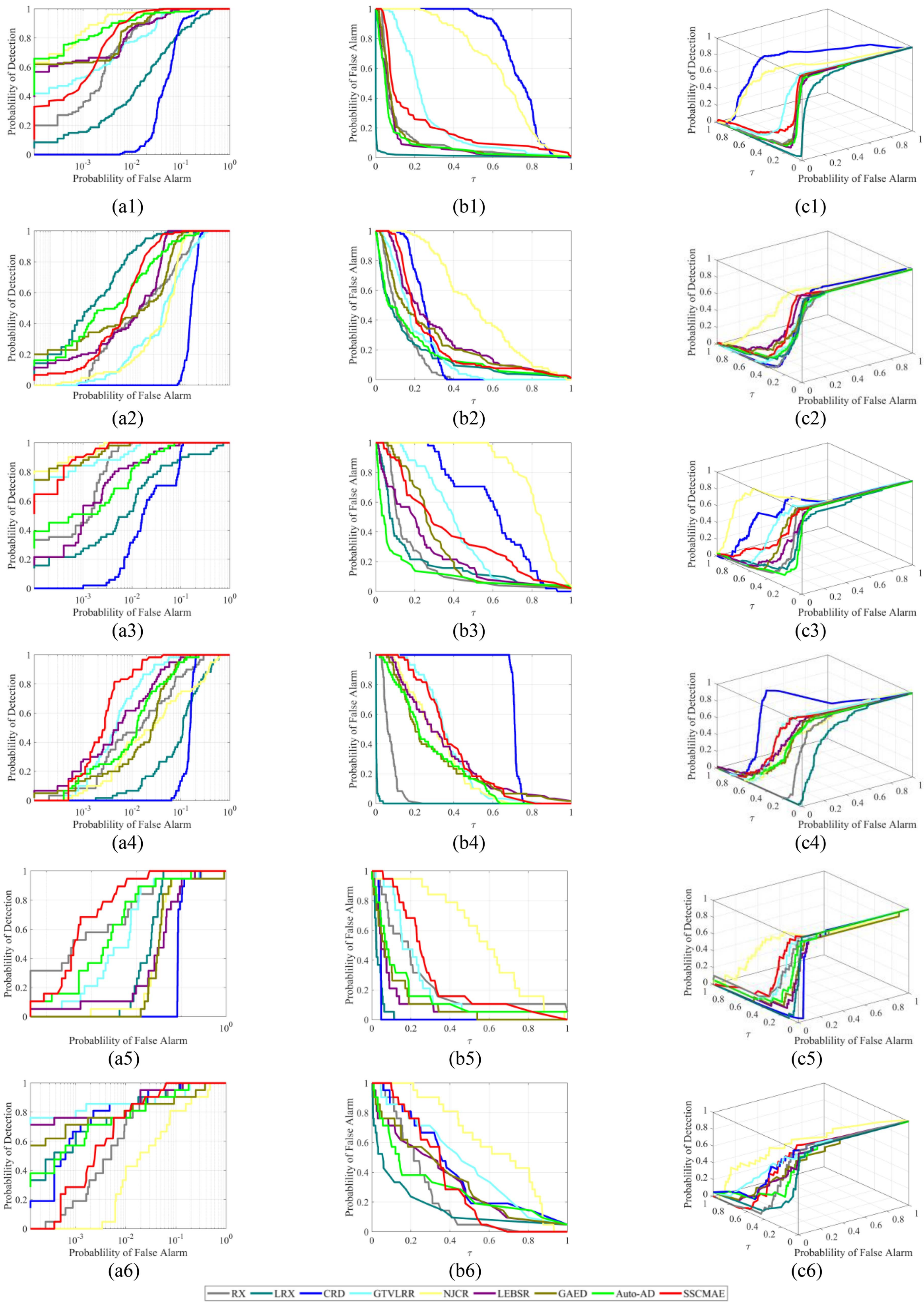
TABLE III
AUC VALUES OF THE COMPARED METHODS ON EACH DATASET

	RX	LRX	CRD	GTVLRR	NJCR	LEBSR	Auto-AD	GAED	SSCMAE
Texas Coast	0.9946	0.9463	0.9394	0.9899	0.9935	0.9932	0.9947	0.9959	**0.9975**
Belcher Bay	0.9622	0.9192	0.8555	0.9385	0.9905	0.9854	0.9835	0.9770	**0.9919**
PaviaC	0.9984	0.9423	0.9633	0.9987	0.9804	0.9928	0.9925	0.9995	**0.9997**
Gulfport	0.9526	0.8494	0.8516	0.9889	0.9833	0.9821	0.9731	0.9683	**0.9954**
Cat Island	0.9807	0.9789	0.9018	0.9813	0.9474	0.9522	0.9871	0.9223	**0.9980**
HYDICE	0.9857	0.9890	0.9908	0.9817	0.9524	0.9905	0.9833	0.9643	**0.9912**

Fig. 7 illustrates the separability of all methods in terms of anomalies and background. Across the six datasets, it is observable that the red anomaly boxes of SSCMAE are not consistently positioned at their highest level. However, the blue background boxes appear narrow, especially noticeable in the Gulfport dataset. This observation suggests the effectiveness of the proposed method in successfully suppressing background information. Furthermore, in comparison to other methods, the proposed approach demonstrates a greater distance and less overlap between the red and blue boxes, indicating its stronger ability to differentiate between targets and background, as well as its good generalization across the six datasets. In conclusion, the SSCAE model, as proposed in this method, demonstrates superior performance over both conventional and state-of-the-art models in terms of anomalous target detection.

E. Ablation Study

1) Effectiveness of Essential Components: In order to assess the effectiveness of each essential element in our proposed SSCMAE, an ablation study was conducted on six datasets. These studies were specifically aimed at investigating the influence of spatial attention, spectral attention, and 3DCAE. Table IV



(a1) (b1) (c1)

(a2) (b2) (c2)

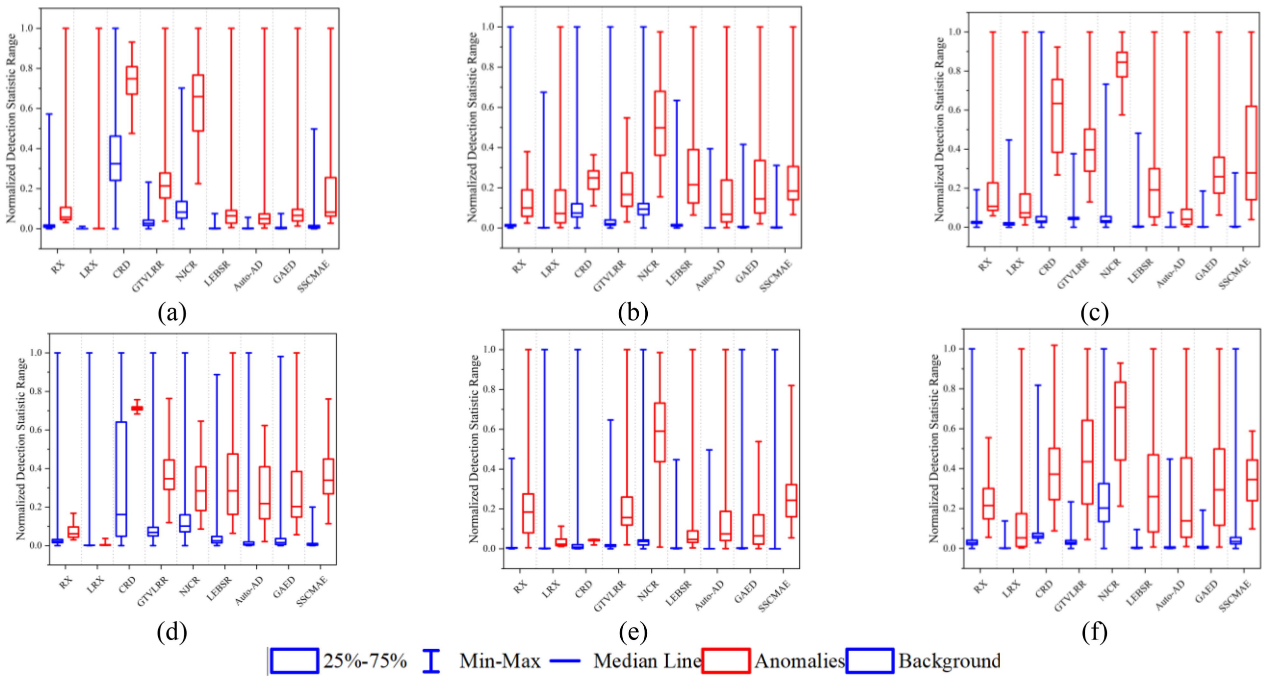
(a3) (b3) (c3)

(a4) (b4) (c4)

(a5) (b5) (c5)

(a6) (b6) (c6)

Fig. 6. ROC curves of the nine methods for the six real datasets. (a) 2-D ROC curves of (PD, PF). (b) 2-D ROC curves of (τ, PF). (c) 3-D ROC curves. (a1)–(c1)–(a6)–(c6) Results for datasets I–VI, respectively.



(a) (b) (c) (d) (e) (f)

Fig. 7. Separability maps obtained by different detectors on each dataset. (a) Texas coast. (b) Belcher bay. (c) Pavia. (d) Gulfport. (e) Cat Island. (f) HYDICE.

TABLE IV
AUC Values of the Compared Methods on Each Dataset

Spectral attention	Spatial attention	3DCAE	Texas Coast	Belcher Bay	PaviaC	Gulfport	Cat Island	HYDICE
√	×	×	0.9947	0.9668	0.9973	0.8885	0.9812	0.9238
×	√	×	0.9705	0.9541	0.9958	0.8293	0.9706	0.9261
√	√	×	0.9952	0.9760	0.9983	0.9056	0.9924	0.9631
√	√	√	**0.9975**	**0.9901**	**0.9997**	**0.9954**	**0.9980**	**0.9912**

TABLE V
AUC Values of the Mask Strategy on Each Datasets

Guided mask	Random mask	Texas Coast	Belcher Bay	PaviaC	Gulfport	Cat Island	HYDICE
×	×	0.8007	0.9292	0.8740	0.8915	0.9558	0.9056
√	×	0.9924	0.9838	0.9966	0.9794	0.9887	0.9611
×	√	0.8238	0.9475	0.9070	0.9684	0.9738	0.9368
√	√	**0.9963**	**0.9901**	**0.9989**	**0.9917**	**0.9907**	**0.9912**

provides a detailed analysis of the relationship between these various components and their corresponding AUC values. When using only spatial attention, the AUC values for the Texas Coast, Belcher Bay, Pavia C, Gulfport, and Cat Island datasets were 0.9705, 0.9541, 0.9958, 0.8293, 0.9706, and 0.9261, respectively. Except for the HYDICE dataset, the AUC values of all six datasets are higher when using spatial information. This is because the HYDICE data contain numerous water absorption bands with low signal-to-noise ratios and poor quality, which affects the extraction of spectral features. This further underscores the crucial role of spectral information and the importance of spatial information in hyperspectral anomaly detection. The incorporation of joint spatial and spectral detection further enhances anomaly detection performance, emphasizing the critical role of the dual-branch structure in SSCMAE. SSCMAE demonstrates its ability to capture intricate spatial and spectral features simultaneously and effectively. In addition, in order to achieve a balance between global and local features, local feature extraction is performed using 3DCAE. The detection performance of 3DCAE on the six datasets shows additional improvement, validating the effectiveness of integrating a lightweight 3DCAE to complement local features.

The proposed algorithm has significantly improved detection accuracy across the Belcher Bay, Gulfport, Cat Island, and HYDICE datasets, with notable enhancements observed, particularly in the Gulfport dataset. This improvement can be attributed to the unique characteristics of these three datasets, which exhibit visual anomalies and more closely resemble their surrounding backgrounds. For example, in the Gulfport dataset, the aircraft is nearly imperceptible to the unaided eye in the pseudocolor image. In contrast, the PaviaC and Texas Coast datasets contain smaller targets, simpler backgrounds, and more noticeable distinctions between visual anomalies and backgrounds. Consequently, anomalies in these datasets are easier to detect with minimal impact from the two extraction methods, namely spectral and spatial. The experimental findings highlight the algorithm's strong and consistent performance in situations where both targets and anomalies exhibit visual resemblance.

2) Effectiveness of Mask Strategy: To evaluate the effectiveness of the mask strategy proposed in our SSCMAE, ablation studies were conducted on six datasets. These studies were specifically aimed at investigating the influence of guided masks and random masks. Table V provides a detailed analysis of the relationship between various masking methods and their corresponding AUC values.

The findings presented in Table V demonstrate that the utilization of the masking strategy consistently improves the detection efficacy of the proposed SSCMAE.

V. Conclusion

This article proposes an SSCMAE for hyperspectral anomaly detection. The SSCMAE consists of a spatial branch and a spectral branch. During the reconstruction process, a specific guided mask is designed to mitigate the emergence of anomalous targets. The generation of this mask incorporates consideration of spectral differences between these anomalies and the background. The interaction between the spatial and spectral branches is achieved through cross-connection convolutional layers, which enhance the spatial and spectral feature representation of the HSI during background reconstruction. In addition, a lightweight 3DCAE is integrated to extract local features, addressing the challenge of ViT's limited effectiveness in learning local structures. The final detection results are determined by calculating the reconstruction error using the Mahalanobis distance. Empirical findings using real-world data illustrate the efficacy of the algorithm and the benefits of the hybrid architecture, which integrates the transformer and 3DCAE for anomaly detection. Our future endeavors will focus on refining ViT to minimize false alarms and enhance the efficiency of the proposed SSCMAE method.

Acknowledgment

The authors would like to thank the relevant researchers for their datasets and codes.

References

[1] F. Xiong, J. Zhou, S. Tao, J. Lu, and Y. Qian, "SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510816.

[2] Y. Zhang, L. Zhang, R. Song, C. Huang, and Q. Tong, "Considering nonoverlapped bands construction: A general dictionary learning framework for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505215.

[3] D. Manolakis, "Taxonomy of detection algorithms for hyperspectral imaging applications," *Opt. Eng.*, vol. 44, no. 6, 2005, Art. no. 066403.

[4] N. Huyan, X. Zhang, H. Zhou, and L. Jiao, "Hyperspectral anomaly detection via background and potential anomaly dictionaries construction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2263–2276, Apr. 2019.

[5] R. Tao, X. Zhao, W. Li, H.-C. Li, and Q. Du, "Hyperspectral anomaly detection by fractional Fourier entropy," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4920–4929, Dec. 2019.

[6] D. Manolakis, E. Truslow, M. Pieper, T. Cooley, and M. Brueggeman, "Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 24–33, Jan. 2014.

[7] J. Liu, Y. Feng, W. Liu, D. Orlando, and H. Li, "Training data assisted anomaly detection of multi-pixel targets in hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 68, pp. 3022–3032, 2020.

[8] M. Bajić, "Airborne hyperspectral surveillance of the ship-based oil pollution in croatian part of the adriatic sea," *Geodetski List*, vol. 66, no. 2, pp. 77–100, 2012.

[9] H. Su, Z. Wu, H. Zhang, and Q. Du, "Hyperspectral anomaly detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 64–90, Mar. 2022.

[10] I. S. 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.

[11] J. M. Molero, E. M. Garzon, I. Garcia, and A. Plaza, "Analysis and optimizations of global and local versions of the RX algorithm for anomaly detection in hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 801–814, Apr. 2013.

[12] H. Kwon and N. M. 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.

[13] 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.

[14] 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.

[15] 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.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[17] A. Ghiasi et al., "What do vision transformers learn? A visual exploration," 2022, *arXiv:221206727*.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[19] C. Zhao, C. Li, S. Feng, and N. Su, "Spectral-spatial stacked autoencoders based on the bilateral filter for hyperspectral anomaly detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2209–2212.

[20] S. Wang, X. Wang, L. Zhang, and Y. Zhong, "Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5503314.

[21] Z. He, D. He, M. Xiao, A. Lou, and G. Lai, "Convolutional transformer-inspired autoencoder for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5508905.

[22] L. Wang, X. Wang, A. Vizziello, and P. Gamba, "RSAAE: Residual self-attention-based autoencoder for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510614.

[23] J. Wang, J. Sun, Y. Xia, and Y. Zhang, "Dynamic negative sampling autoencoder for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9829–9841, 2022.

[24] W. Xie, T. Jiang, Y. Li, X. Jia, and J. Lei, "Structure tensor and guided filtering-based algorithm for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4218–4230, Jul. 2019.

[25] B. Tu, Z. Wang, H. Ouyang, X. Yang, J. Li, and A. Plaza, "Hyperspectral anomaly detection using the spectral–spatial graph," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542814.

[26] A. Ng, "Sparse autoencoder," *CS294A Lect. Notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[27] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:201011929*.

[28] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[29] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–296.

[30] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7242–7252.

[31] J. Wang, Y. Liu, and L. Li, "Background augmentation with transformer-based autoencoder for hyperspectral anomaly detection," in *Proc. Int. Conf. Intell. Sci.*, 2022, pp. 302–309.

[32] S. Xiao, T. Zhang, Z. Xu, J. Qu, S. Hou, and W. Dong, "Anomaly detection of hyperspectral images based on transformer with spatial–spectral dual-window mask," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1414–1426, 2023.

[33] Y. Zhang, Y. Fan, and M. Xu, "A background-purification-based framework for anomaly target detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1238–1242, Jul. 2020.

[34] K. Yang, H. Sun, C. Zou, and X. Lu, "Cross-attention spectral–spatial network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518714.

[35] 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, Oct. 2019.

[36] N. Huyan, X. Zhang, D. Quan, J. Chanussot, and L. Jiao, "Cluster-memory augmented deep autoencoder via optimal transportation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531916.

[37] F. Verdoja and M. Grangetto, "Graph Laplacian for image anomaly detection," *Mach. Vis. Appl.*, vol. 31, no. 1/2, 2020, Art. no. 11.

[38] X. Kang, X. Zhang, S. Li, K. Li, J. Li, and J. A. Benediktsson, "Hyperspectral anomaly detection with attribute and edge-preserving filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5600–5611, Oct. 2017.

[39] W. Xie, J. Lei, B. Liu, Y. Li, and X. Jia, "Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection," *Neural Netw.*, vol. 119, pp. 222–234, 2019.

[40] X. Wang, Y. Wang, Z. Mu, and M. Wang, "FCAE-AD: Full convolutional autoencoder based on attention gate for hyperspectral anomaly detection," *Remote Sens.*, vol. 15, no. 17, 2023, Art. no. 4263.

[41] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Glob. Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, 2008.

[42] T. Cheng and B. Wang, "Graph and total variation regularized low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 391–406, Jan. 2020.

[43] T. Guo, L. He, F. Luo, X. Gong, L. Zhang, and X. Gao, "Learnable background endmember with subspace representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5501513, doi: 10.1109/TGRS.2023.3341245.

[44] S. Chang and P. Ghamisi, "Nonnegative-constrained joint collaborative representation with union dictionary for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5534913, doi: 10.1109/TGRS.2022.3195339.

[45] P. Xiang, S. Ali, S. K. Jung, and H. Zhou, "Hyperspectral anomaly detection with guided autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5538818.

Qing Guo received the B.E. degree in earth information science and technology from the China University of Geosciences, Wuhan, China, in 2022. She is currently working toward the M.Sc. degree in cartography and geographic information system with the National Engineering Laboratory for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her major research interests include hyperspectral target detection and unsupervised learning.

Yi Cen received the B.E. degree in GIS and cartography from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2001, and the M.S. and Ph.D. degrees in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2004 and 2008, respectively.

She is currently an Associate Professor with the Hyperspectral Remote Sensing Division, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. Her main research interests include hyperspectral imagery, machine learning, and target detection.

Lifu Zhang (Senior Member, IEEE) received the B.E. degree in photogrammetry and remote sensing from the Department of Airborne Photogrammetry and Remote Sensing and the M.E. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, both from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1992 and 2000, respectively, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, in 2005.

He is currently a full-time Professor and the Dean with the Hyperspectral Remote Sensing Division, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include hyperspectral remote sensing and imaging spectrometer system development and its applications.

Dr. Zhang is a Member of the SPIE, the Academy of Space Science of China, and the Chinese National Committee of the International Society for Digital Earth (CNISDE), a Vice-Chairman of the Hyperspectral Earth Observation Committee, CNISDE, and a Standing Committeeman of the Expert Committee of China Association of Remote Sensing Applications.

Yan Zhang received the B.E. degree in remote sensing science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2017, and the M.S. degree in science and technology of surveying and mapping from the China University of Petroleum (east China), Qingdao, China, in 2020. She is currently working toward the Ph.D. degree in cartography and geographic information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her current research interests include hyperspectral and multispectral image fusion.

Yixiang Huang received the B.E. degree in software engineering from the Information Science and Technology College, Dalian Maritime University, Dalian, China, in 2020. He is currently working toward the Ph.D. degree in cartography and geographic information system with the National Engineering Laboratory for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interest is related to hyperspectral change detection.