





Anomaly Detection of Hyperspectral Images Based on Transformer With Spatial–Spectral Dual-Window Mask

Song Xiao , Member, IEEE, Tian Zhang , Zhangchun Xu, Jiahui Qu , Member, IEEE, Shaoxiong Hou, Graduate Student Member, IEEE, and Wenqian Dong , Member, IEEE

Abstract—Anomaly detection has become one of the crucial tasks in hyperspectral images processing. However, most deep learning-based anomaly detection methods often suffer from the incapability of utilizing spatial–spectral information, which decreases the detection accuracy. To address this problem, we propose a novel hyperspectral anomaly detection method with a spatial–spectral dual-window mask transformer, termed as S2DWMTrans, which can fully extract features from global and local perspectives, and suppress the reconstruction of anomaly targets adaptively. Specifically, the dual-window mask transformer aggregates background information of the entire image from a global perspective to neutralize anomalies, and uses neighboring pixels in a dual-window to suppress anomaly reconstruction. An adaptive-weighted loss function is designed to further suppress anomaly reconstruction adaptively during network training process. According to our investigation, this is the first work to apply transformer to hyperspectral anomaly detection. Comparative experiments and ablation studies demonstrate that the proposed S2DWMTrans achieves competitive performance.

Index Terms—Anomaly detection, dual-window mask transformer (DWMTrans), hyperspectral image (HSI).

I. INTRODUCTION

HYPERSPECTRAL image (HSI) with rich spectral information has a powerful ability of distinguishing different materials, and have been widely used in various remote sensing

data analysis applications [1], [2], such as anomaly detection [3], [4], image classification [5], [6], and target detection [7], [8]. Among these applications, anomaly detection utilizes continuous spectral information and spatial information of land covers to detect anomalies with significant different spectral signatures from their surrounding environment [9].

Anomaly detection has the unique advantage of not requiring prior spectral information, which prompts a lot of detection methods to be proposed [10], [11]. According to the separation standards of background and anomaly, hyperspectral anomaly detection methods can be divided into three categories: statistical model-based, reconstruction model-based, and deep learning-based. The statistical model-based anomaly detection methods can be traced back to the RX method proposed by Reed et al. [12], which assumes that the background part can be represented by the Gaussian background distribution model. However, the HSI background may be very complex and cannot be fitted by the Gaussian model only. In order to improve the detection performance, local RX (LRX) [13], weighted RX (WRX) [14], and other methods were proposed. LRX selects a specific area as the background, usually in the form of sliding dual windows. However, it is difficult to determine the size of the dual window. WRX can better estimate the background by assigning low weights to possible anomalous pixels and high weights to other pixels. Zhang et al. proposed an isolation forest method for hyperspectral anomaly detection, which is based on Ostu and assume that the isolation of the abnormal pixel from the alternative pixel is more sensitive [15]. Sertac et al. designed an anomaly detection algorithm for HSIs, which is based on nonparametric Bayesian background estimation [16]. Chang et al. designed an orthogonal subspace projection target detector for hyperspectral anomaly detection, which takes advantage of automatic target generation process, and achieved an outstanding performance [17]. With the development of the compressed sensing theory and machine learning, anomaly detection method based on reconstruction has gradually become a research hotspot. The basic idea is to use an overcomplete dictionary to represent the HSI, and the difference between the reconstructed image and the original image is used as the basis for anomaly judgment. Li et al. [18] proposed an anomaly detection method based on background joint sparse representation. The joint sparse model constructs a sample set that can represent the background distribution, and then, performs sparse representation of the

Manuscript received 26 October 2022; revised 10 December 2022 and 22 December 2022; accepted 23 December 2022. Date of publication 4 January 2023; date of current version 24 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101414; in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 095920221320; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2021JQ-194 and Grant 2021JQ-197; in part by the Fundamental Research Funds for the Central Universities under Grant XJS210108 and Grant XJS210104; in part by the China Post-Doctoral Science Foundation under Grant 2021M702546 and Grant 2021M702548; in part by the Scientific and Technological Activities for Overseas Students of Shaanxi Province under Grant 2020-017; and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515110856. (Corresponding author: Song Xiao.)

Song Xiao is with the Department of Electronic and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China, and also with the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: xs_xidian@163.com).

Tian Zhang, Zhangchun Xu, Jiahui Qu, Shaoxiong Hou, and Wenqian Dong are with the State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China (e-mail: ztnwu0111@163.com; xzchxx@163.com; jhqu@xidian.edu.cn; sxhou@stu.xidian.edu.cn; wqdong@xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3232762

test pixel based on the sample set. The degree of difference between the test pixel and the original pixel reflects the anomaly degree of the test pixel. Li et al. [19] described an collaborative representation-based detector (CRD), which adopts the local dual-window strategy, and uses the pixels between the two windows to carry out collaborative representation of the central pixels.

In recent years, deep learning has achieved excellent results in the fields of computer vision and remote sensing [20], [21]. For example, Xie et al. designed a spectral–spatial anomaly detection method for HSI, which is based on band selection. Specifically, they fully utilized the underlying physical characteristics to train the unsupervised network [22]. Mihai et al. proposed a deep convolutional model to detect anomalies in burned area contexts of Sentinel-2 scenes, which adopt a self-supervised paradigm to learn the image representations [23]. The convolution neural network (CNN) is the typical representative of deep learning, which needs sufficient samples for supervised training [24], [25]. However, hyperspectral anomaly detection does not have any prior information and the data samples are limited. In order to solve this shortcoming, Li et al. [26] combined the CNN with transfer learning to realize anomaly detection. Compared with the CNN, which requires reference images and a large number of data samples, autoencoder (AE), generative adversarial net (GAN) [27], and adversarial auto encoder (AAE) [28] are more suitable for HSI anomaly detection. Zhang et al. [29] constructed an adaptive subspace model based on stacked AEs to extract deep features with differences for anomaly detection. Manifold learning is adopted in literature [30] to extract local spatial information and integrate it with AE. Xie et al. [31] introduced spectral constraint into AAE to suppress background while maintaining abnormal characteristic information, making it easier to distinguish anomalies from background.

At present, the traditional methods and deep learning-based methods both have obtained satisfied detection performance. However, most of the current research only consider spectral information or local spatial information. The traditional anomaly detection methods have some limitations in extracting image features manually, which may ignore a lot of details. Among the deep-learning-based methods, those improved AE occupy the majority and are usually trained by spectral vectors, so the exploration of spatial information is insufficient. The anomaly detection methods based on the CNN can only obtain local spatial information.

To overcome these limitations, this article proposes a hyperspectral anomaly detection method named spatial–spectral dual-window mask transformer (S2DWMTrans) to make full use of the spatial–spectral information of the HSI. First, convolution operation is introduced to give the model local receptive fields, and then, the dual perspective spatial–spectral feature extraction and fusion module (DPS2FEFM) is used to extract spatial–spectral feature under two perspectives. In the global perspective, the background information with a high proportion is gathered. In the local perspective, the neighbor features in the dual window are refined and fused to reconstruct the image. In addition, due to the relatively large reconstruction error of

anomalies in the early training stage, the S2DWMTrans designed an adaptive weighted loss function (AWLF) to give a small weight to the pixels with large reconstruction error to reduce the contribution to the total loss. So, the network not only realizes the full extraction of spatial–spectral information, but also effectively suppresses anomalies and achieves accurate background reconstruction. In order to further improve the detection accuracy, the input and reconstructed images are postprocessed. The error of these two images is used as the weight, and a nonlinear mapping is introduced to enlarge the background and anomaly in the original image. Finally, the Mahalanobis distance detector is used to get the detection result.

The main contributions of this article are as follows.

- 1) A novel S2DWMTrans is proposed to fully utilize the spatial–spectral information in HSIs for competitive anomaly detection performance. Specifically, a dual-window mask transformer (DWMTrans) is contained in the proposed S2DWMTrans, which can fully exploit the combined spatial–spectral features of the HSI in the global and local perspectives for reconstruction, and exert different degrees of suppression on anomalies respectively.
- 2) An AWLF is designed to further suppress anomaly reconstruction by reducing the weight of potential anomalies during the network training process.
- 3) We conduct several comparative experiments and ablation studies on five datasets to demonstrate the advancement and effectiveness of our proposed S2DWMTrans. The detection performance outperforms many advanced methods on all the datasets.

II. RELATED WORK

The transformer was first proposed by Google researchers [32] and applied in the field of natural language processing, which is similar to AE in overall structure. A transformer consists of encoder and decoder, and requires input data in vector form. Dosovitskiy et al. [33] proposed the vision transformer model (ViT), which removes the decoder and only uses the encoder structure. In order to meet the strict requirements of the format of the transformer input data, the ViT has a preprocessing process that flattens the input image into vectors. The transformer despite the traditional CNN structure, whose whole network architecture is completely composed of attention. To be precise, the transformer only consists of self-attention module and feed forward neural network (FFN). The self-attention module is the core component of the transformer [34], and has been applied to the many image recognition task and its specific structure is shown in Fig. 1 [35], [36].

The input data \mathbf{X} are first transformed into Query (Q), Key (K), and Value (V) matrices through three linear transformations, which can be expressed as

$$\begin{aligned} \mathbf{Q} &= \mathbf{XW}^q \\ \mathbf{K} &= \mathbf{XW}^k \\ \mathbf{V} &= \mathbf{XW}^v \end{aligned} \quad (1)$$

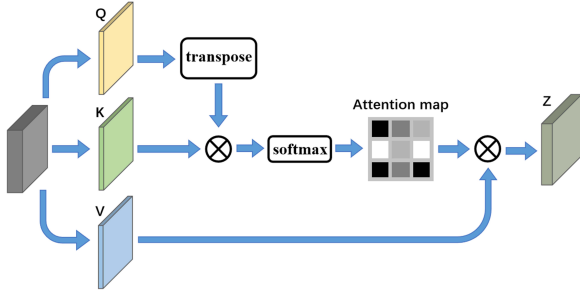


Fig. 1. Flowchart of the self-attention module.

where \mathbf{W}^q , \mathbf{W}^k , and \mathbf{W}^v denote the weight matrices of linear transformations for generating the query, key, and value tensors, respectively, which are trainable and can improve the model fit ability. Each line of \mathbf{Q} , \mathbf{K} , and \mathbf{V} represents a sample data. Multiplying \mathbf{Q}^T and \mathbf{K} can calculate the similarity between the extracted features \mathbf{Q} and \mathbf{K} , which indicates the similarity between the features of each pixel and other pixels. The inner product between each row vector of matrices \mathbf{Q} and \mathbf{K} is calculated to get the weight matrix, which shows how much attention each pixel in the input gets when the model focuses on a particular pixel. Then, the weight matrix is divided by d_k to prevent the inner product from being too large, and the normalized weight matrix is obtained by softmax function. The complete calculation formula can be expressed as

$$\mathbf{Z} = \text{softmax} \left(\frac{\mathbf{Q}^T \mathbf{K}}{\sqrt{d_k}} \right) \mathbf{V} \quad (2)$$

where d_k is the sample coding dimension.

III. METHODOLOGY

A. Architecture Overview

Just as with any other anomaly detection tasks, how to well represent the background samples while poorly perform the anomaly samples is the key to clearly distinguish small anomaly samples. In this article, we propose a hyperspectral anomaly detection method named S2DWMTrans to make full use of the spatial–spectral information to suppress the anomaly reconstruction.

The overall architecture of the proposed S2DWMTrans is represented in Fig. 2, which mainly consists of three parts, which are as follows.

- 1) A DWMTrans is constructed for fully digging the spatial–spectral feature in global and local perspectives for reconstruction, and applying different degrees of inhibitory effect on the anomaly samples.
- 2) An AWLF is designed to further suppress anomaly reconstruction by reducing the weights of potential anomalies during network training.
- 3) A postprocessing module is applied in order to further improve the detection accuracy.

B. Dual-Window Mask Transformer (DWMTrans)

A DWMTrans is constructed for fully extracting spatial–spectral features from both global and local perspectives. A local

shallow feature extraction module (LSFEM) is first designed to initially extract local features and inject adjacent information into the spectral vectors, and thus, enhance the information interaction between image pixels and surrounding areas. Next, the spectral vectors are fed into the DPS2FEFM, which consists of several cascaded dual-window mask encoder block (DWMEB), to fuse the spatial–spectral information from two perspectives to jointly suppress anomaly reconstruction. Specifically, the multi-head self-attention (MSA) in DWMEB with a global receptive field is used to gather background information of the whole image to neutralize anomalies. In the local view, the information under the dual-window mask are mined by the constructed dual-window mask multihead self-attention (DWM-MSA) in DWMEB to suppress anomaly reconstruction. Finally, the refined features are passed through the background reconstruction module (BRM) to accurately learn the background distribution function.

1) *Local Shallow Feature Extraction Module (LSFEM)*: In order to enhance the information interaction between pixels and surrounding areas, the input HSI image of the DWMTrans is first passed through an LSFEM. As shown in Fig. 2, the conv layer of the LSFEM can traverse the entire image to extract local features, and gather the feature information in the convolution kernel area to the center pixel. The following batch normalization (BN) accelerates the speed of the network training and prevents overfitting. Then, a rectified liner unit (ReLU) layer is followed to learn a nonlinear representation. Since HSIs have multiple spectral bands and rich spectral information, the spectral vector of a single pixel can act as an input vector of the transformer. Given the HSI patch $\mathbf{H} \in \mathbb{R}^{M \times N \times B}$ with $M \times N$ pixels and B spectral bands as the input of the proposed S2DWMTrans, which can also be regarded as $M \times N$ vectors with B dimensions. The aforementioned process of the LSFEM can be more intuitively formulated as

$$\mathbf{Y} = f_{\text{LSFEM}}(\mathbf{H}) = f_{\text{conv}}(\mathbf{H}) \quad (3)$$

$$\mathbf{y} = \text{Flatten}(\mathbf{Y}) = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \quad (4)$$

where \mathbf{Y} represents the output of the LSFEM, \mathbf{y} represents the flattened vectors as input to the next module. As the input of the next module, all the spectral vectors passing through the LSFEM carry the local spatial information in the vicinity, and thus, fuse the local and global information.

2) *Dual Perspective Spatial–Spectral Feature Extraction and Fusion Module (DPS2FEFM)*: Since most regions in HSIs are background, it is necessary to combine global features of background to weaken anomaly reconstruction using background information. At the same time, considering that the spectral characteristics of abnormal pixels are different from those of neighboring pixels, it is necessary to obtain detailed information of neighboring regions to represent abnormal pixels and suppress anomaly reconstruction.

We design a DPS2FEFM that can obtain the global and local spatial–spectral information simultaneously. As shown in Fig. 2, the DPS2FEFM is a stack of the DWMEB, whose calculation formula is as follows:

$$F_{\text{DPS2FEFM}} = F_{\text{DWMEB}, N} \dots F_{\text{DWMEB}, i} \dots F_{\text{DWMEB}, 1}[\mathbf{y}] \quad (5)$$

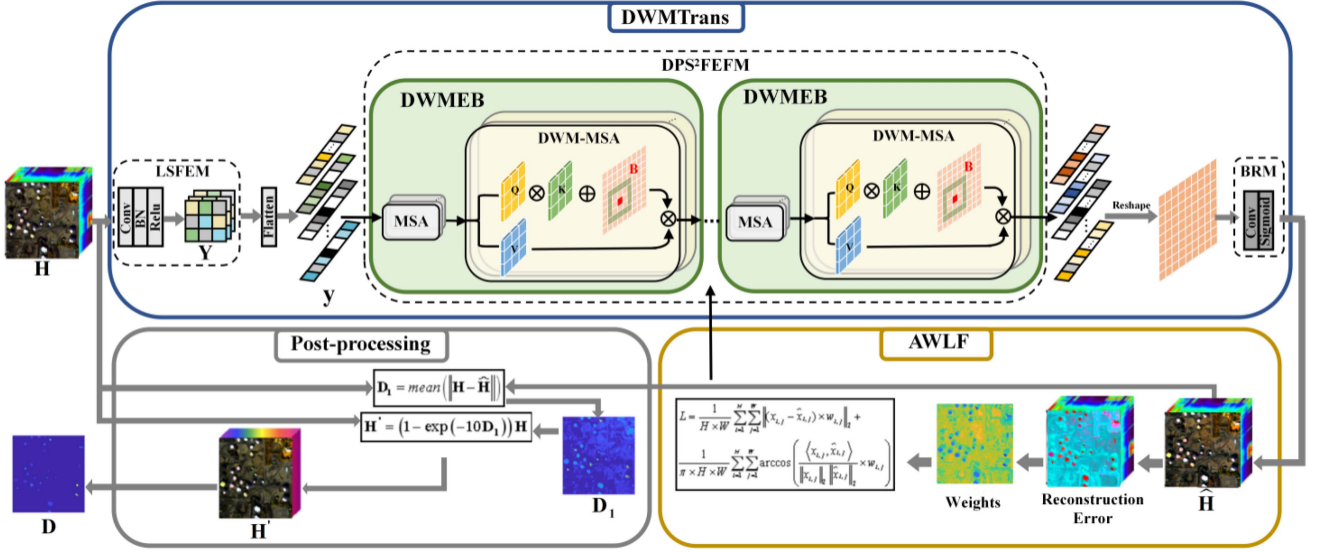


Fig. 2. Architecture of the proposed S2DWMTrans.

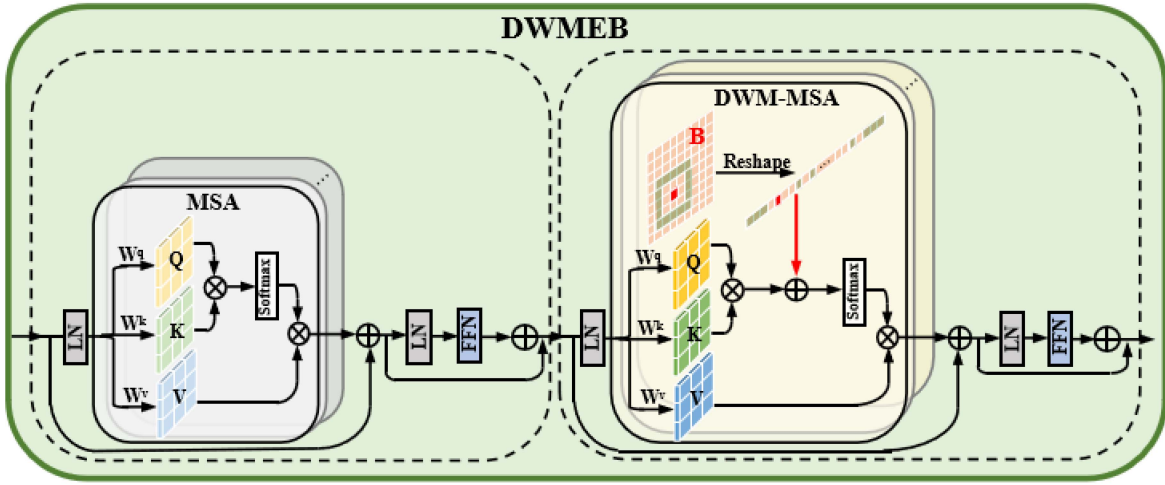


Fig. 3. Flowchart of the the DWMEB.

where $F_{\text{DWMEB},i}[\cdot]$ represents the function of the i th DWMEB. Fig. 3 shows the specific structure of the DWMEB, which consists of two consecutive encoder-based structures. The first one is the same as transformer encoder in the ViT and is used to extract global features. The second one designs a DWM-MSA, which limits the calculation of the attention weight of self-attention to a dual-window region and is used to extract local features. The function of the DWMEB can be expressed as

$$F_{\text{DWMEB},i} = F_{\text{DWM-MSA},i} [F_{\text{MSA},i} (\mathbf{o}_{i-1})] \quad (6)$$

where $F_{\text{MSA},i}(\cdot)$ represents the function of the first encoder, $F_{\text{DWM-MSA},i}(\cdot)$ represents the function of the DWM-MSA-based structure, and \mathbf{o}_{i-1} represents the output of i -1th DWMEB. $F_{\text{MSA},i}(\cdot)$ is replaced by \mathbf{m} for the sake of representation

$$\begin{aligned} \mathbf{m} &= F_{\text{MSA},i}(\mathbf{o}_{i-1}) = \mathbf{o}_{i-1} + f_{\text{MSA},i}(\text{LN}(\mathbf{o}_{i-1})) \\ &\quad + \text{FFN}[\text{LN}[\mathbf{o}_{i-1} + f_{\text{MSA},i}(\text{LN}(\mathbf{o}_{i-1}))]] \end{aligned} \quad (7)$$

where $f_{\text{MSA},i}(\cdot)$ represents the function of the MSA block. $F_{\text{DWM-MSA},i}(\cdot)$ can be described in detail as follows:

$$\begin{aligned} F_{\text{DWM-MSA},i} &= \mathbf{m} + f_{\text{DWM-MSA},i}(\text{LN}(\mathbf{m})) \\ &\quad + \text{FFN}[\text{LN}[\mathbf{m} + f_{\text{DWM-MSA},i}(\text{LN}(\mathbf{m}))]] \end{aligned} \quad (8)$$

where $f_{\text{DWM-MSA},i}(\cdot)$ represents the function of the DWM-MSA block.

The MSA is same as the multihead self-attention module in the ViT, whose formula is as follows:

$$\begin{cases} \mathbf{a}^g = \sum_{i=1}^{N_a^g} w_i^{ag} \mathbf{a}_i^g + \sum_{i=1}^{N_b^g} w_i^{bg} \mathbf{b}_i^g \\ \mathbf{b}^g = \sum_{i=1}^{N_a^g} w_i^{ag} \mathbf{a}_i^g + \sum_{i=1}^{N_b^g} w_i^{bg} \mathbf{b}_i^g \end{cases} \quad (9)$$

where \mathbf{a}^g and \mathbf{b}^g are vectors of abnormal and background pixels, respectively. g is short for global. w_i^{ag} is the attention weights of pixels to be reconstructed and abnormal pixels. w_i^{bg} is the

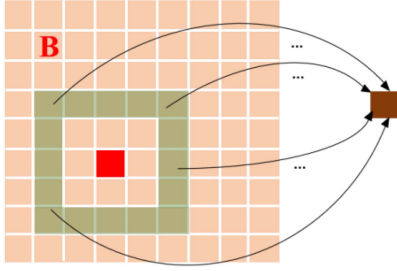


Fig. 4. Flowchart of the local attention-based reconstruction.

attention weights of pixels to be reconstructed and background pixels. N_a^g and N_b^g are the total number of abnormal and background pixels in the corresponding region, respectively, which remain unchanged. If the pixel to be reconstructed is abnormal, w_i^{bg} is far less than w_i^{ag} . However, N_b^g is much larger than N_a^g , which makes the majority of background pixels affect the reconstruction of anomalies. If the pixel to be reconstructed is background, most of the background pixels can be well represented, and only a few small background regions deviating from the overall background area will be considered as abnormal pixels.

DWM-MSA only calculates the attention weight of all pixels between the inner and outer windows and the central pixels, that is, adding a dual-window attention mask B . Fig. 4 shows the schematic of B , and its calculation formula is as follows:

$$\mathbf{Z} = \mathbf{W}\mathbf{V} = \text{softmax} \left[\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B} \right] \mathbf{V} \quad (10)$$

where W represents the attention weight between pixels to be reconstructed and other pixels, and Z is the output of DWM-MSA. The detailed calculation formula of DWM-MSA can be expressed as

$$\begin{aligned} \mathbf{a}^l &= \sum_{i=1}^{N_a^l} w_i^{al} \mathbf{a}_i^l + \sum_{i=1}^{N_b^l} w_i^{bl} \mathbf{b}_i^l \\ \mathbf{b}^l &= \sum_{i=1}^{N_a^l} w_i^{al} \mathbf{a}_i^l + \sum_{i=1}^{N_b^l} w_i^{bl} \mathbf{b}_i^l \end{aligned} \quad (11)$$

where l is short for local. N_a^l and N_b^l will change with the size of dual window mask.

As the pixel to be reconstructed may be background or anomaly, formula (9) will correspond to the following four situations as shown in Fig. 5.

- 1) As shown in Fig. 5(a), when the pixel to be reconstructed is abnormal and N_b^l is much larger than N_a^l , similar to global reconstruction, the background pixels will inhibit the reconstruction of abnormal pixel.
- 2) As shown in Fig. 5(b), the pixel to be reconstructed is abnormal and N_b^l is close to N_a^l . This situation is very rare and can be solved by adjusting the size of the inner and outer windows.
- 3) As shown in Fig. 5(c), when the pixel to be reconstructed is background and N_a^l is zero, the pixel will be completely reconstructed.

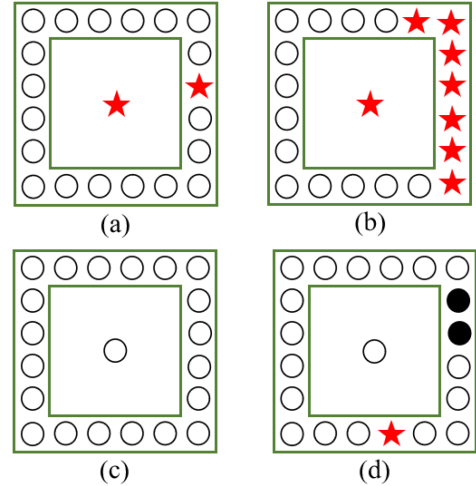


Fig. 5. Diagram of four different reconstruction. The red star represents the abnormal pixel, and the black and white circles represent the background pixel with different distributions.

- 4) As shown in Fig. 5(d), the pixel to be reconstructed is background, and N_a^l is the total number of background pixels with different distributions and abnormal pixels.

In the previous MSA, a small number of background pixels with a different distribution than the one to be reconstructed were considered anomalies. This situation will be improved in the DWM-MSA. Since most of pixels in the dual window are background pixels with similar distribution and w^{bl} is much larger than w^{al} , the small background area will slowly return to its own distribution. MSA and DWM-MSA appear in pairs in the DWMEB, which fully integrate global and local features, effectively suppressing anomalies and highlighting the background.

3) *Background Reconstruction Module (BRM)*: The BRM reconstructs the deep features extracted previously. The module first uses a conv layer to adjust the number of spectral bands of the image. The subsequent sigmoid function is used to limit the value of the output image between 0 and 1, which is convenient to calculate the loss with the input image.

C. Adaptive-Weighted Loss Function (AWLF)

Compared with the background, anomalies account for a small proportion in HSIs and usually exist in irregular small areas, which are difficult to reconstruct. However, the conv layers and DWMEB with strong learning ability have made the total loss in the training process decrease continuously. As the data fitting degree of the network gradually increased, anomalies will gradually be reconstructed. In order to curb the training trend and further suppress the anomalies, the reconstruction errors in the training process can be used to accurately suppress the potential anomalies.

Since anomaly pixels have a large reconstruction error in the early stage of network training, the AWLF is designed to reduce the weight of areas with large reconstruction errors and increase the weight of other areas when calculating loss. This function helps suppress the reconstruction of anomalies and highlights the reconstruction of the background. The reconstruction error

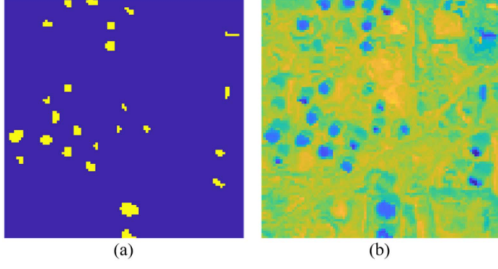


Fig. 6. (a) Distribution of anomalies. (b) Distribution of weights.

calculation formula of a single pixel can be expressed as

$$e_{i,j} = \|\mathbf{x}_{i,j} - \hat{\mathbf{x}}_{i,j}\|_2 \quad (12)$$

where $\mathbf{x}_{i,j}$ and $\hat{\mathbf{x}}_{i,j}$ are the spectral vectors of the input image and reconstructed image at position (i,j) , respectively. Further, the weight of the pixel can be expressed as

$$w_{i,j} = \frac{\left(\text{Norm}\left(\frac{1}{e_{i,j}}\right) + 1 - \exp(-10 \times e_{i,j})\right)}{2} \quad (13)$$

where $\text{Norm}(\cdot)$ represents normalization. The distribution of pixels and their corresponding weights is shown in the Fig. 6. Potential anomalies [yellow areas in Fig. 6(a)] will be given smaller weights [darker blue areas in Fig. 6(b)].

In the training process of the DWMTrans, $w_{i,j}$ is updated every 20 epochs. During the first 20 epochs, $w_{i,j}$ is initialized to 1, and its value would not change after five times of updating. The adaptive-weighted loss function L based on $w_{i,j}$ can be expressed as

$$L = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \|\mathbf{x}_{i,j} - \hat{\mathbf{x}}_{i,j}\|_2 \times w_{i,j} + \frac{1}{\pi \times H \times W} \sum_{i=1}^H \sum_{j=1}^W \arccos\left(\frac{\langle \mathbf{x}_{i,j}, \hat{\mathbf{x}}_{i,j} \rangle}{\|\mathbf{x}_{i,j}\|_2 \|\hat{\mathbf{x}}_{i,j}\|_2} \times w_{i,j}\right) \quad (14)$$

where H and W are the height and width of the image, respectively, and $\langle \mathbf{x}_{i,j}, \hat{\mathbf{x}}_{i,j} \rangle$ represents the inner product of two spectral vectors. The first term of L uses the mean square error (MSE) to quantify spatial differences between the input image and the reconstructed image, while the second term uses spectral angle mapper (SAM) to constrain the spectral similarity between images. Each term is multiplied by the weight $w_{i,j}$ when calculating the error at a single pixel. Because the weight corresponding to the abnormal pixel is small, its loss contributes little to the total loss, and thus, the abnormal pixel avoids being reconstructed by the network.

D. Postprocessing

The proposed DWMTrans obtains the network mapping function F_{DWMTrans} after completing the iterative training process. The input HSI \mathbf{H} is passed through F_{DWMTrans} and obtain the

abnormal inhibited reconstructed image $\hat{\mathbf{H}}$, which can be expressed as

$$\hat{\mathbf{H}} = F_{\text{DWMTrans}}(\mathbf{H}). \quad (15)$$

In order to improve detection accuracy, postprocessing of input and reconstructed images is carried out as shown in Fig. 2. The postprocessing module first calculates the reconstruction error between the two images and averages it on the channel dimension, so as to obtain the preliminary anomaly detection result \mathbf{D}_1 . \mathbf{D}_1 can be expressed as,

$$\omega = 1 - \exp(-10\mathbf{D}_1) \quad (16)$$

where ω ranges from 0 to 1. The reconstruction error corresponding to the abnormal pixel is relatively large, so its weight is close to 1, while the reconstruction error corresponding to the background pixel is very small, so its weight is close to 0. Then, the weight matrix ω is multiplied by the input \mathbf{H} to increase the separability of background and anomaly. The obtained background suppressed image can be expressed as

$$\mathbf{H}' = (1 - \exp(-10\mathbf{D}_1)) \mathbf{H}. \quad (17)$$

After this operation, the abnormal pixels with large \mathbf{D}_1 will maintain the spectral vector in \mathbf{H} , while the spectral vector of background pixels with small \mathbf{D}_1 will be suppressed. Finally, the obtained \mathbf{H}' has great separability in anomaly and background, and anomaly detection can be realized by calculating the Mahalanobis distance on \mathbf{H}' .

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets and Experimental Setup

In this article, five real hyperspectral datasets are used to verify the effectiveness of the proposed S2DWMTrans¹ [37], whose pseudocolor and reference images are shown in Fig. 7. Gulfport dataset was taken by Airborne Visible/Infrared Imaging Spectrometer Sensor (AVIRIS) in 2010 over Gulfport, USA, with the ground spatial resolution of 3.4 m. The size of Gulfport dataset and its ground truth is 100×100 . Three aircraft contained in the image features were taken as abnormal targets, which occupies 155 60 pixels. After the contaminated bands are removed, 191 spectral bands are left, covering a spectral range from 400 to 2500 nm.

Pavia dataset was taken by the reflective optics system imaging spectrometer sensor (ROSIS) in Pavia, Italy. The image include rivers, bridges, soil, and buildings. The resolution of the image is 150×150 , with 175 spectral bands and a wavelength range of 430 to 860 nm. The ground spatial resolution is 1.3 m. The size of the ground truth of Pavia dataset is 150×150 , in which some vehicles on the bridge occupying a total of 68 pixels are considered as abnormal targets.

Texas Coast dataset was acquired by an AVIRIS sensor in August 2010 over the coast of Texas, USA, with a spatial resolution of 17.2 m. The image contains 207 bands, each with a size of 100×100 , so does its ground truth. There is a parking lot

¹<http://xudongkang.weebly.com/>

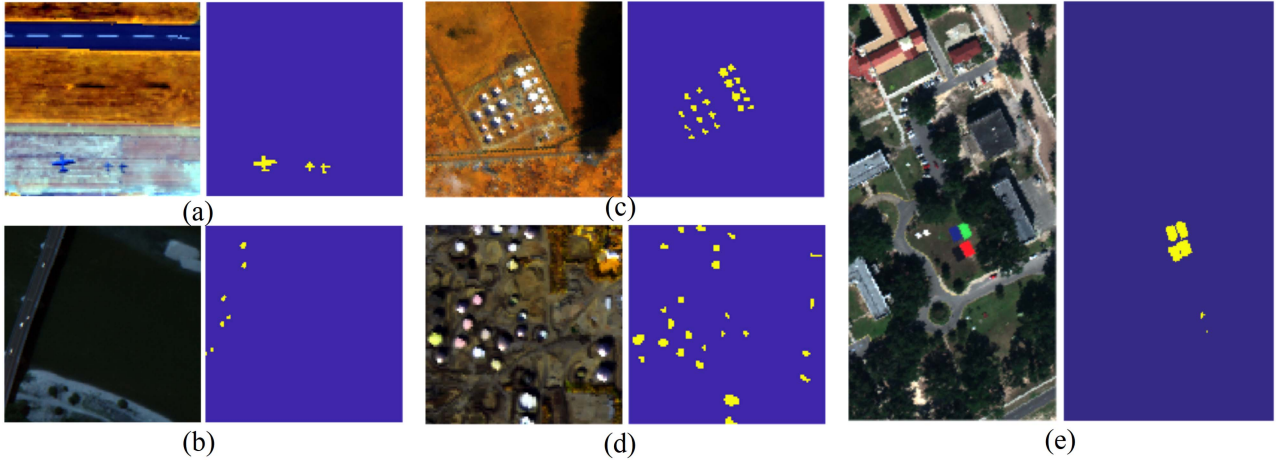


Fig. 7. Pseudocolor diagram and reference diagram of (a) Gulfport dataset, (b) Pavia dataset, (c) Texas Coast dataset, (d) Los Angeles dataset, and (e) Muufl dataset.

in the ground object scene, in which some vehicles occupying a total of 272 pixels are marked as abnormal targets.

Los Angeles dataset was acquired by the AVIRIS sensor in November 2011 in Los Angeles, USA, with a spatial resolution of 7.1 m. The image scene and its ground truth both cover 100×100 pixels and has 205 spectral bands, among which some noise bands have been removed. As for Los Angeles dataset, some storage tanks occupying a total of 272 pixels are considered as targets.

Muufl dataset was acquired by the ITRES Research Ltd. (ITRES) Compact Airborne Spectrographic Imager (CASI)-1500 sensor in 2010, covering the University of Southern Mississippi Gulfport Campus, Mississippi, USA.² The image contains 64 bands, covering the wavelength range of 367.7–1043.4 nm. The size of Gulfport dataset and its ground truth is 220×325 , and the spatial resolution is 0.54×1.0 m. In this article, we crop a subimage of 140×280 pixels from the upper right of the whole image, in which four cloths occupying a total of 269 pixels are considered as targets.

Some of the anomalies in the five experimental datasets are in the form of points, some show small irregular regions, and some have obvious structural information. In addition, their spatial resolution, number of spectral bands and ground object scenes are mostly different. Therefore, experiments on these datasets can effectively verify the effectiveness and robustness of the proposed method.

The proposed DWMTrans is implemented with the PyTorch 1.7 and Python 3.7 on Ubuntu, and trained on the GeForce GRX 3090 GPU. The postprocessing is implemented with the MATLAB 2018a on Windows10, and trained on the CPU Intel(R) Core(TM) i7-9750H. DWMTrans uses Adam [38] optimization to find the optimal solution, and the learning rate is set as 0.0001. Each batch of sample size is the number of pixels in the whole image, and the number of training epochs is 400. The number of output nodes of the LSFEM is 128, the dimension of the

spectral vector is set to 64, and the number of nodes of the two full connection layers in the FFN is 256 and 128, respectively.

B. Quality Assessment

There are qualitative and quantitatively indicators to evaluate HSI anomaly detection. One of these indicators is receiver operating characteristic (ROC), which is determined by the probability of detection P_d and false alarm rate P_f of abnormal detection results. P_d and P_f can be expressed as

$$P_d = \frac{N_d}{N_t} \quad (18)$$

$$P_f = \frac{N_f}{N_{\text{total}}} \quad (19)$$

where N_d , N_t , N_f , and N_{total} represent the number of detected abnormal pixels, the number of abnormal pixels in the reference image, the number of background pixels mistaken as abnormal pixels in the detection result, the total number of pixels in HSIs, respectively. Taking P_f as abscissa and P_d as ordinate, the corresponding ROC curve can be drawn. The closer the ROC curve is to the upper left corner, the better the performance of the method. Another metric is area under curve (AUC), namely, the area under the ROC curve, which can be expressed as

$$\text{AUC} = \int_0^1 \text{ROC}(x) dx. \quad (20)$$

The higher the AUC value is, the better the performance of the corresponding algorithm is. Precision-recall curve (PRC) is also an indicator we used. The closer the PRC curve is to the upper right corner, the better the performance of the model.

C. Ablation Study

1) *Parameters Analysis*: In order to make the detection result of the proposed method as optimal as possible, multiple experiments are conducted to select relatively optimal parameters.

a) *Sizes of w_{in} and w_{out}* : The most important factors influencing the choice of w_{in} and w_{out} is the size of the

²<https://datasets.bifrost.ai/info/1773>

TABLE I
AUC OF DIFFERENT PARAMETER COMBINATIONS FOR PAVIA DATASET

| win_{in} | win_{out} | 7 | 9 | 11 | 13 |
|------------|-------------|--------|--------|--------|--------|
| 3 | --- | 0.9964 | 0.9963 | 0.9963 | 0.9965 |
| 5 | --- | 0.9965 | 0.9964 | 0.9964 | 0.9963 |
| 7 | --- | --- | 0.9964 | 0.9964 | 0.9965 |
| 9 | --- | --- | --- | 0.9965 | 0.9963 |

TABLE II
AUC OF DIFFERENT PARAMETER COMBINATIONS FOR TEXAS COAST DATASET

| win_{in} | win_{out} | 7 | 9 | 11 | 13 |
|------------|-------------|--------|--------|--------|--------|
| 3 | --- | 0.9992 | 0.9993 | 0.9994 | 0.9993 |
| 5 | --- | 0.9991 | 0.9993 | 0.9993 | 0.9978 |
| 7 | --- | --- | 0.9992 | 0.9994 | 0.9992 |
| 9 | --- | --- | --- | 0.9976 | 0.9991 |

TABLE III
AUC OF DIFFERENT PARAMETER COMBINATIONS FOR LOS ANGELES DATASET

| win_{in} | win_{out} | 7 | 9 | 11 | 13 |
|------------|-------------|--------|--------|--------|--------|
| 3 | --- | 0.9929 | 0.9932 | 0.9958 | 0.9964 |
| 5 | --- | 0.9925 | 0.9935 | 0.9965 | 0.9966 |
| 7 | --- | --- | 0.9925 | 0.9963 | 0.9959 |
| 9 | --- | --- | --- | 0.9953 | 0.9930 |

TABLE IV
AUC OF DIFFERENT PARAMETER COMBINATIONS FOR GULFPORT DATASET

| win_{in} | win_{out} | 7 | 9 | 11 | 13 |
|------------|-------------|--------|--------|--------|--------|
| 3 | --- | 0.9809 | 0.9830 | 0.9834 | 0.9829 |
| 5 | --- | 0.9831 | 0.9826 | 0.9831 | 0.9831 |
| 7 | --- | --- | 0.9830 | 0.9833 | 0.9869 |
| 9 | --- | --- | --- | 0.9833 | 0.9835 |

TABLE V
AUC OF DIFFERENT PARAMETER COMBINATIONS FOR MUUFL DATASET

| win_{in} | win_{out} | 7 | 9 | 11 | 13 |
|------------|-------------|--------|--------|--------|---------------|
| 3 | --- | 0.9970 | 0.9965 | 0.9972 | 0.9972 |
| 5 | --- | --- | 0.9965 | 0.9870 | 0.9938 |
| 7 | --- | --- | 0.9957 | 0.9960 | 0.9970 |
| 9 | --- | --- | --- | 0.9972 | 0.9965 |

anomaly target contained in the image. The value range of win_{in} and win_{out} is set as [3,5,7,9] and [7,9,11,13], respectively. Considering $win_{in} < win_{out}$, there are 13 different combinations of window parameters. To determine the optimal settings for win_{in} and win_{out} , we performed parameters analysis on all five datasets for different parameter combinations, and AUC values are used for judging indicators. Tables I–V show AUC values of different parameter combinations for Pavia dataset, Texas Coast dataset, Los Angeles dataset, Gulfport dataset, and MUUFL dataset, respectively. From all the experimental results, it can be

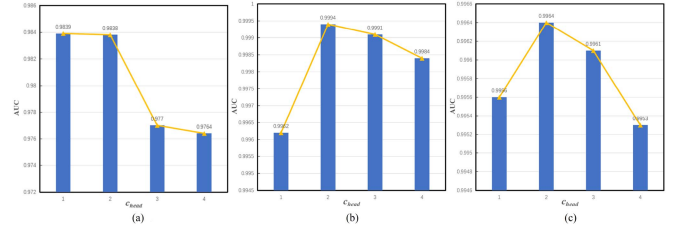


Fig. 8. AUC of different c_{head} on three datasets. (a) Gulfport. (b) Texas Coast. (c) Los Angeles.

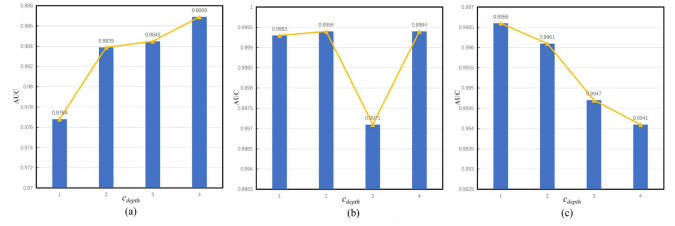


Fig. 9. AUC of different c_{depth} on three datasets. (a) Gulfport. (b) Texas Coast. (c) Los Angeles.

seen that when $win_{in}=3,5,7$ and $win_{out}=13$, the value of AUC is the optimal.

b) Number of MSA and DWM-MSA c_{head} : MSA and DWM-MSA appear in pairs in the DWMEB, which fully integrates global and local features, effectively suppressing anomalies and highlighting the background. If the number of MSA and DWM-MSA c_{head} is too small, the network model will not pay enough attention to the image information. If c_{head} is too large, the calculation will increase cost exponentially. So, c_{head} is set from 1 to 4. In order to determine the optimal c_{head} , we experimented with different values of c_{head} on three datasets. AUC values are used for judging indicators. Fig. 8 shows AUC values of different c_{head} on Gulfport dataset, Texas Coast dataset, and Los Angeles dataset, respectively.

c) Number of DWMEB c_{depth} : The DWMEB can fully excavate the spatial-spectral features of the image from the global and local perspectives for reconstruction, and exert different degrees of suppression on the anomalies. The number of the DWMEB c_{depth} is a key parameter that determines the performance of the network. If the network is too shallow, it is not enough to extract abstract semantic features. If the network is too deep, it is easy to cause overfitting. As DWMEB contains two transformer encoder, the value range of c_{depth} is set from 1 to 4. Fig. 9 shows AUC values of different c_{depth} on three datasets, from which can be seen that when $c_{depth} = 1$ or 2, the value of AUC has a high probability of obtaining the optimal value.

2) Component Analysis: In order to verify the effectiveness of some modules or components in the proposed S2DWMTrans, multiple ablation experiments are carried out. The experimental results of all the five datasets of ablation experiments are shown in Fig. 10, where Base represents the proposed S2DWMTrans without any changes, Base-LSFEM represents that the LSFEM module is removed, Base (MSA) represents that the DWM-MSA is replaced by MSA, Base-AWLF represents that the use of unweighted loss functions for back propagation of the network.

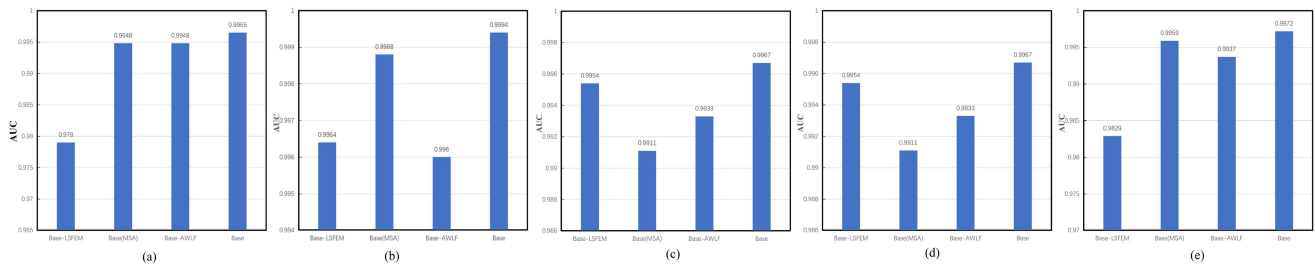


Fig. 10. AUC of different variants of the proposed S2DWMTrans on five datasets. (a) Pavia. (b) Texas Coast. (c) Los Angeles. (d) Gulfport. (e) MUUFL.

a) *Effectiveness of the LSFEM*: The main function of LSFEM is to obtain the spectral vectors with local spatial information, and serve as the input of subsequent transformer to promote the fusion of local and global information. The variable condition of the first ablation experiment is whether the LSFEM exists. As we can see from Fig. 10, the use of LSFEM has resulted in a significant increase in AUC values on all datasets. For example, it can be seen from Fig. 10(a) and (b) that the use of LSFEM increases AUC by 1.75% and 0.3% on Pavia dataset and Texas Coast dataset, respectively, which shows the advantages of LSFEM.

b) *Effectiveness of the DWM-MSA*: The DWMEB consists of two consecutive transformer encoder, of which the second encoder adopts DWM-MSA to limit the calculation of attention weight to a region between dual window, so as to obtain more detailed local information. The variable condition of the second ablation experiment is to replace all DWM-MSA in the network with ordinary MSA. As can be seen from Fig. 10, compared with the model without DWMEB, the proposed model achieves higher AUC on all datasets. For example, the use of the DWMEB has resulted in a significant increase by 0.56% on Los Angeles dataset, which shows the advantages of DWM-MSA.

c) *Effectiveness of the AWLF*: The AWLF assigns a small weight to potential abnormal pixels with bigger reconstruction error, so as to reduce the contribution of anomalies to the total loss, and thus, achieve the purpose of inhibiting abnormal reconstruction. The variable condition of the third ablation experiment is whether different weights are assigned to image pixels during the loss calculation. As shown in Fig. 10, the comparative experimental results between Base-AWLF and the proposed model prove the importance of AWLF for improving AUC. As we can see from Fig. 10, the result of Base is always the best on five datasets. The addition of three modules or components will give positive feedback to the network, which verifies the effectiveness and robustness of the proposed S2DWMTrans.

D. Experimental Result

In order to effectively evaluate the detection performance of the proposed S2DWMTrans, five advanced HSI anomaly detection methods are selected for comparison, including RX [14], CRD [19], LRSMD [39], AED [37], and Auto-AD method [40]. RX is a pioneering work in the field of hyperspectral anomaly detection, which makes use of the statistical characteristics of

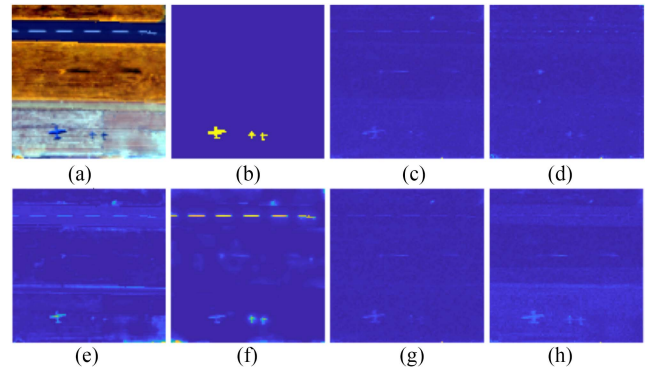


Fig. 11. Detection results of different methods on Gulfport dataset. (a) Pseudocolor Image. (b) Reference. (c) RX. (d) CRD. (e) LRSMD. (f) AED. (g) Auto-AD. (h) S2DWMTrans.

image data to judge the pixels. CRD is a typical detection method based on reconstruction. LRSMD is a low-rank and sparse matrix factorization-based method. AED realizes detection by filtering and differential operations on spatial attributes. Auto-AD uses a CNN with skip connection to reconstruct the background and suppress anomalies at the same time. The reconstruction error of each pixel is calculated to obtain the anomaly detection graph.

The detection results of Gulfport dataset are shown in Fig. 11, in which CRD almost fails to correctly detect abnormal targets. LRSMD and AED detect the traffic dotted line on the highway. LRSMD fails to detect the two small planes, while AED regards the dotted line on the highway as the most likely abnormal target, leading to obvious false detection. RX and Auto-AD only detect larger aircraft targets, while the two small aircraft are faintly visible.

Fig. 12 shows the detection results on Pavia dataset. The detection effect of RX, CRD, and LRSMD is not obvious. AED mistakenly detects the large area beach and long strip bridge in the image as abnormal targets. Auto-AD also detects the edge of the beach and bridge, which may be due to the obvious feature transformation at the junction of various scenes.

The detection results on Texas Coast dataset are shown in Fig. 13. Only a few abnormal targets are detected by CRD. RX and AED also miss some abnormal targets, and AED misdetects in the lower left and right corner of the image where there were no abnormal targets. Although LRSMD and Auto-AD have detected almost all abnormal targets, their confidence of

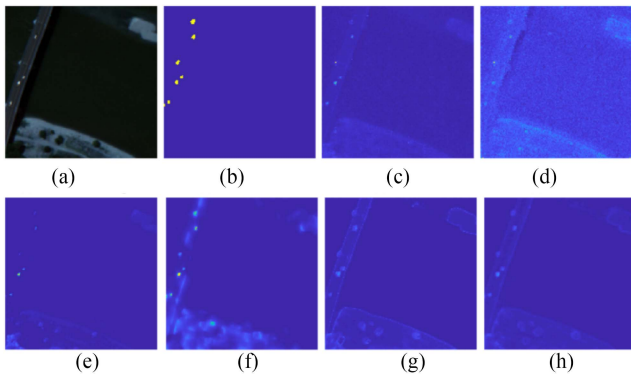


Fig. 12. Detection results of different methods on Pavia dataset. (a) Pseudocolor Image. (b) Reference. (c) RX. (d) CRD. (e) LRSMD. (f) AED. (g) Auto-AD. (h) S2DWMTrans.

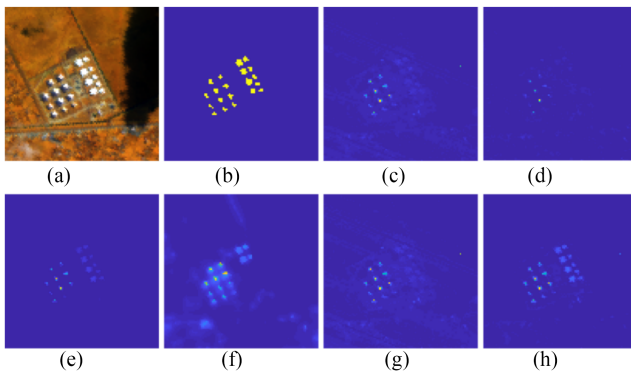


Fig. 13. Detection results of different methods on Texas Coast dataset. (a) Pseudocolor Image. (b) Reference. (c) RX. (d) CRD. (e) LRSMD. (f) AED. (g) Auto-AD. (h) S2DWMTrans.

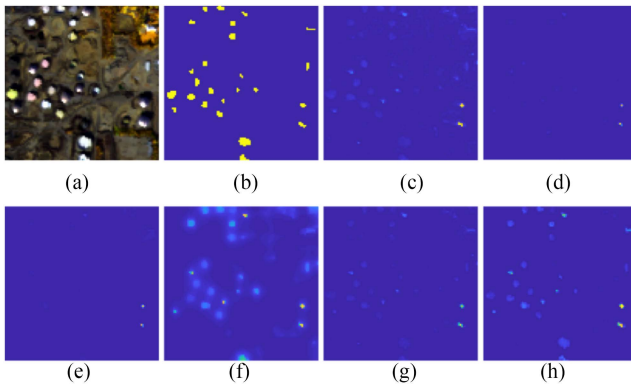


Fig. 14. Detection results of different methods on Los Angeles dataset. (a) Pseudocolor Image. (b) Reference. (c) RX. (d) CRD. (e) LRSMD. (f) AED. (g) Auto-AD. (h) S2DWMTrans.

some abnormal target regions is relatively lower than that of the proposed S2DWMTrans.

The detection results on Los Angeles dataset are shown in Fig. 14. Among them, CRD and LRSMD only detect two obvious abnormal targets. The abnormal targets detected by AED are very fuzzy, resulting in misjudgment at the connection between

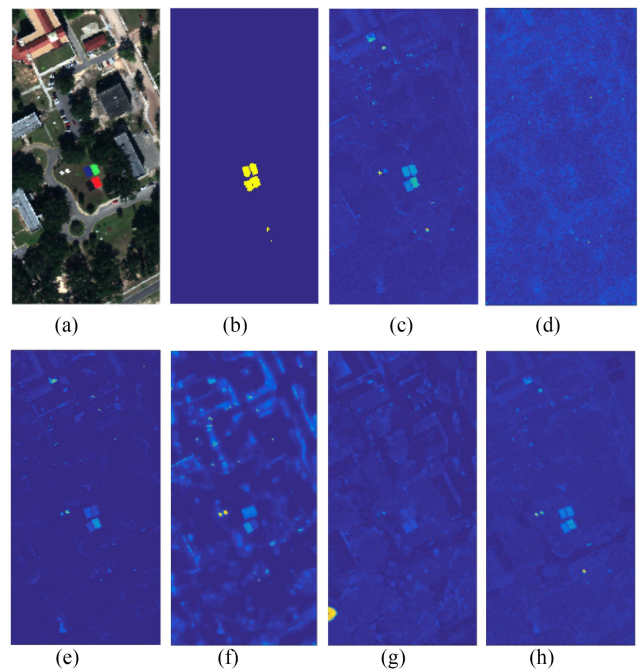


Fig. 15. Detection results of different methods on MUUFL dataset. (a) Pseudocolor Image. (b) Reference. (c) RX. (d) CRD. (e) LRSMD. (f) AED. (g) Auto-AD. (h) S2DWMTrans.

TABLE VI
AUC VALUES OF DIFFERENT METHODS

| | RX | CRD | LRSMD | AED | Auto-AD | S2DWMTrans |
|-------------|--------|---------------|---------------|---------------|---------|---------------|
| Gulfport | 0.9521 | <u>0.6391</u> | 0.8781 | 0.9314 | 0.9797 | 0.9967 |
| Pavia | 0.9534 | 0.9009 | <u>0.8358</u> | 0.9793 | 0.9905 | 0.9965 |
| Texas Coast | 0.9546 | 0.9548 | 0.9975 | <u>0.7958</u> | 0.9958 | 0.9994 |
| Los Angeles | 0.9887 | 0.9475 | 0.9852 | <u>0.8397</u> | 0.9908 | 0.9967 |
| MUUFL | 0.9962 | <u>0.6700</u> | 0.9864 | 0.9384 | 0.9565 | 0.9972 |

abnormal targets. RX and Auto-AD have good detection results, but there is also a small area of error detection in the upper right corner of the image.

Fig. 15 shows the detection results on MUUFL dataset. It can be obviously observed that the proposed method obtains excellent performance. It can be seen that CRD hardly works on this image. As for other competitors, AED and Auto-AD both have high false alarm rate, which is easy to judge the background as abnormal. RX and LRSMD all achieve quite nice results. The proposed S2DWMTrans can suppress the reconstruction of background more effectively so as to highlight the anomalous target with the least false detection. According to the results of the four datasets, the S2DWMTrans detects more abnormal targets and suppresses most of the background. The S2DWMTrans can effectively achieve the separation of background and abnormal target, and has a low rate of missed and false detection.

Both qualitative and quantitative indicators are integrated to jointly evaluate the proposed method. Table VI lists the corresponding AUC values on five datasets, where the values

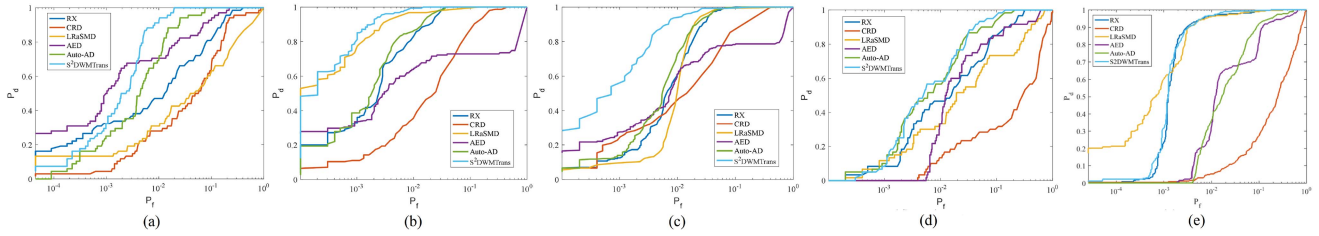


Fig. 16. ROC curves of different methods. (a) Pavia. (b) Texas Coast. (c) Los Angeles. (d) Gulfport. (e) MUUFL.

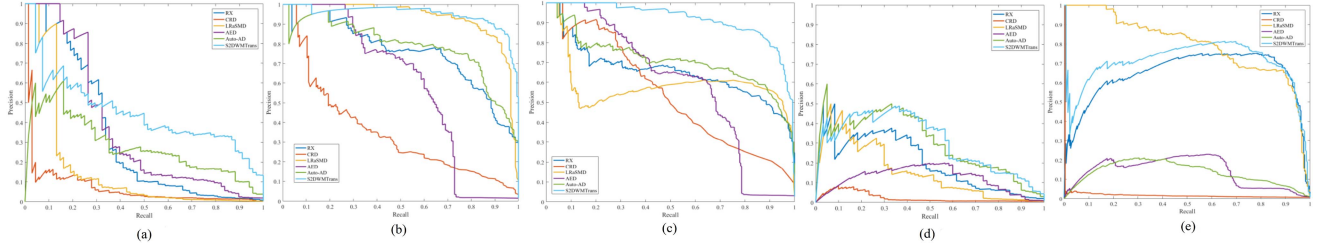


Fig. 17. PRC curves of different methods. (a) Pavia. (b) Texas Coast. (c) Los Angeles. (d) Gulfport. (e) MUUFL.

in bold and underlined represent the optimal and worst results. It can be seen from the table that the proposed S2DWMTrans has the highest AUC value on all datasets, which means that the S2DWMTrans has the best detection ability on the selected datasets. As can be seen from Fig. 16, the ROC curves of the S2DWMTrans are almost always higher than those of other methods. Combined with the two quantitative indexes, the S2DWMTrans has the best detection effect on the five datasets. Fig. 17 shows the PRC curves of different methods on five datasets. Compared with other methods, it can be observed that the PRC curves of the proposed S2DWMTrans is closer to the upper right corner, which reflects that the S2DWMTrans achieves higher detection performance. In Fig. 17(c), the PRC curve produced by the S2DWMTrans is consistently at the top, indicating that our method is optimal on the Los Angeles dataset. The area under the PRC curves of the S2DWMTrans for the all five datasets are almost bigger than those of RX, CRD, LRaSMD, AED, and Auto-AD, which indicates that the proposed S2DWMTrans has satisfactory target detection and background suppression capabilities.

V. CONCLUSION

In this article, we present a new hyperspectral anomaly detection method based on the S2DWMTrans to solve the problem that the existing methods do not make full use of the spatial-spectral information of HSIs. The proposed method fully extracts spatial-spectral joint features from global and local perspectives by using the constructed DWMTrans, and further reconstructs them from these two perspectives. In the global perspective, all background information is integrated to weaken the abnormal features. In the local perspective, the neighbor information is used to greatly constrain the abnormal features, so as to effectively inhibit the reconstruction of abnormal targets. In order to further reconstruct image background and suppress anomalies, an AWLF is proposed to suppress potential abnormal

reconstruction precisely. The experiments of the S2DWMTrans are carried out on five datasets to evaluate the effectiveness and robustness of the proposed method.

REFERENCES

- [1] H. Su, Y. Yu, Z. Wu, and Q. Du, "Random subspace-based k -nearest class collaborative representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6840–6853, Aug. 2021.
- [2] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Trans. Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.
- [3] Y. Li, J. Wang, X. Liu, N. Xian, and C. Xie "DIM moving target detection using spatio-temporal anomaly detection for hyperspectral image sequences," in *Proc. IGARSS/IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 7086–7089, doi: [10.1109/IGARSS.2018.8517601](https://doi.org/10.1109/IGARSS.2018.8517601).
- [4] C.-I. Chang, "Target-to-anomaly conversion for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–28, 2022, Art. no. 5540428, doi: [10.1109/TGRS.2022.3211696](https://doi.org/10.1109/TGRS.2022.3211696).
- [5] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.
- [6] W. Dong, T. Zhang, J. Qu, S. Xiao, T. Zhang, and Y. Li, "Multibranch feature fusion network with self- and cross-guided attention for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, Jun. 7, 2022, Art. no. 5530612.
- [7] B. Du, L. Zhang, D. Tao, and D. Zhang, "Unsupervised transfer learning for target detection from hyperspectral images," *Neurocomputing*, vol. 1, no. 59, pp. 72–82, 2013.
- [8] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58–69, Jan. 2002.
- [9] K. Jiang, W. Xie, Y. Li, J. Lei, G. He, and Q. Du, "Semisupervised spectral learning with generative adversarial network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 5224–5236, Jul. 2020.
- [10] S. Li, K. Zhang, P. Duan, and X. Kang, "Hyperspectral anomaly detection with kernel isolation forest," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 319–329, Jan. 2020.
- [11] F. Verdoja and M. Grangetto, "Graph Laplacian for image anomaly detection," *Mach. Vis. Appl.*, vol. 1, no. 59, 2020, Art. no. 11.
- [12] 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. 8, no. 10, pp. 1760–1770, Oct. 1990.

- [13] H. Kwon, S. Z. Der, and N. M. Nasrabadi, "Adaptive anomaly detection using subspace separation for hyperspectral imagery," *Opt. Eng.*, vol. 1, no. 59, pp. 3342–3351, 2003.
- [14] Q. Guo, B. Zhang, and Q. Ran, "Weighted-RXD and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2351–2366, Jun. 2014.
- [15] Y. Zhang, Y. Dong, K. Wu, and T. Chen, "Hyperspectral anomaly detection with Otsu-based isolation forest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9079–9088, 2021.
- [16] S. Arisoy and K. Kayabol, "Nonparametric Bayesian background estimation for hyperspectral anomaly detection," *Digit. Signal Process.*, vol. 111, Apr. 2021, Art. no. 102993.
- [17] C.-I. Chang, H. Cao, and M. Song, "Orthogonal subspace projection target detector for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4915–4932, Mar. 25, 2021.
- [18] J. Li, H. Zhang, L. Zhang, and L. Ma, "Hyperspectral anomaly detection by the use of background joint sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2523–2533, Jun. 2015.
- [19] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.
- [20] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian pyramid dense network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, May 14, 2021, Art. no. 5507113.
- [21] W. Dong, S. Hou, S. Xiao, J. Qu, Q. Du, and Y. Li, "Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7303–7317, Dec. 2021.
- [22] W. Xie, Y. Li, J. Lei, J. Yang, C.-I. Chang, and Z. Li, "Hyperspectral band selection for spectral–spatial anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3426–3436, May 2020.
- [23] M. Coca, I. C. Neagoe, and M. Datcu, "Hybrid DNN-dirichlet anomaly detection and ranking: Case of burned areas discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Sep. 16, 2022, Art. no. 4414116.
- [24] W. Dong, T. Zhang, J. Qu, Y. Li, and H. Xia, "A spatial–spectral dual-optimization model-driven deep network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5542016.
- [25] J. Qu, S. Hou, W. Dong, S. Xiao, Q. Du, and Y. Li, "A dual-branch detail extraction network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Nov. 23, 2021, Art. no. 5518413.
- [26] W. Li, G. Wu, and Q. Du, "Transferred deep learning for hyperspectral target detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5177–5180.
- [27] J. M. P. Goodfellow and I. Pouget-Abadie, "Generative adversarial nets," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, vol. 60, 2014, pp. 2672–2680.
- [28] A. Makhzani, J. Shlens, and N. Jaitly, "Adversarial autoencoders," 2016, *arXiv:1511.05644 [cs.LG]*.
- [29] L. Zhang and B. Cheng, "A stacked autoencoders-based adaptive subspace model for hyperspectral anomaly detection - sciencedirect," *Infrared Phys. Technol.*, vol. 60, pp. 52–60, 2019.
- [30] X. Lu, W. Zhang, and J. Huang, "Exploiting embedding manifold of autoencoders for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1527–1537, Mar. 2020.
- [31] W. Xie, J. Lei, and B. Liu, "Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection," *Neural Netw.*, vol. 60, pp. 222–234, 2019.
- [32] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," 2017, *arXiv:1706.03762 [cs.CL]*.
- [33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.
- [34] W. Hu, Y. Huang, L. Wei, H. Zhang, and F. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, 2021, Art. no. 258619.
- [35] F. Wang et al., "Residual attention network for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458.
- [36] L. Drew, S. Dan, and E. Sven, "Global-and-local attention networks for visual recognition," 2018, *arXiv:abs/1805.08819*.
- [37] 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.
- [38] J. Kingma and D. P. Ba, "Adam: A method for stochastic optimization," 2017.
- [39] 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.
- [40] 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, pp. 1–14, Mar. 22, 2021, Art. no. 5503314.



Song Xiao (Member, IEEE) received the M.S. degree in communication and information system and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2001 and 2004, respectively.

From 2006 to 2007, she was with the Viterbi School of Engineering, University of Southern California, as a Postdoctoral Researcher. She is currently a Professor and Ph.D. Director with Beijing Electronic Science and Technology Institute, Beijing, China, also a Professor and Ph.D. Director with the State Key

Laboratory of Integrated Service Network, Xidian University. She has authored more than 80 international journal and conference papers. Her research interests include image compression and coding, joint source channel coding, multimedia transmission systems over wired/wireless network, and compressed sensing.

Prof. Xiao is the Secretary-General of Image Application in Military and Civilian Integration (IAMCI) Professional Committee of the China Society of Image and Graphics. She is a Council Member of the Shaanxi Society of Image and Graphics and is a Member of the IEEE Multimedia Communication Technology Committee and the IEEE Signal Processing Society.



Tian Zhang received the B.E. degree in communication engineering from Northwest University, Xi'an, China, in 2020. She is currently working toward the M.S. degree in communication engineering with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an.

Her research interests include joint classification of hyperspectral image and LiDAR image, and deep learning.



Zhangchun Xu received the bachelor's degree in communication engineering from Hohai University, Nanjing, China, in 2019. He is currently working toward the master's degree in information and communications engineering with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China.

His research interests include hyperspectral image fusion, anomaly detection, and deep learning.



Jiahui Qu (Member, IEEE) received B.S. degree in communication engineering from Yantai University, Yantai, China, in 2014, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in June 2020.

She was an exchange Ph.D. Student of Mississippi State University from 2018 to 2019, supervised by Dr. Q. Du. She is currently a Lecturer with the State Key Laboratory of Integrated Services Networks, Xidian University. She has authored and co-authored more than 20 papers in known academic journals and

conferences, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the *Remote Sensing*. Her research interests include hyperspectral image detection, image fusion, neural networks, and deep learning.



Shaoxiong Hou (Graduate Student Member, IEEE) received the bachelor's degree in communication engineering from the Xi'an University of Technology, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree in information and communications engineering in information and communications engineering with the State Key Laboratory of Integrated Service Networks of Xidian University, Xi'an.

His research interests include hyperspectral image fusion, hyperspectral anomaly detection, hyperspectral change detection, and deep learning.



Wenqian Dong (Member, IEEE) received B.S. degree in communication engineering from Yantai University, Yantai, China, in 2014, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in June 2020.

She has been a Visiting Scholar with Simon Fraser University, Burnaby, BC, Canada, from 2018 to 2019. She is currently a Lecturer with the State Key Laboratory of Integrated Services Networks, Xidian University. She has authored and co-authored more than 20 papers in refereed journals and conferences, including the *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, the *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, and the *Remote Sensing*. Her research interests include compressed sensing, machine learning, and hyperspectral remote sensing image processing.