An Unsupervised Momentum Contrastive Learning Based Transformer Network for Hyperspectral Target Detection

Yulei Wang, *Member, IEEE*, Xi Chen, Enyu Zhao, *Member, IEEE*, Chunhui Zhao, Meiping Song, and Chunyan Yu

***Abstract*—Hyperspectral target detection plays a pivotal role in various civil and military applications. Although recent advancements in deep learning have largely embraced supervised learning approaches, they often hindered by the limited availability of labeled data. Unsupervised learning, therefore, emerges as a promising alternative, yet its potential has not been fully realized in current methodologies. This article proposes an innovative unsupervised learning framework employing a momentum contrastive learning-based transformer network specifically tailored for hyperspectral target detection. The proposed approach innovatively combines transformer-based encoder and momentum encoder networks to enhance feature extraction capabilities, adeptly capturing both local spectral details and long-range spectral dependencies through the novel overlapping spectral patch embedding and a cross-token feedforward layer. This dual-encoder design significantly improves the model's ability to discern relevant spectral features amidst complex backgrounds. Through unsupervised momentum contrastive learning, a dynamically updated queue of negative sample features is utilized so that the model can demonstrate superior spectral discriminability. This is further bolstered by a unique background suppression mechanism leveraging nonlinear transformations of cosine similarity detection results, with two nonlinearly pull-up operations, significantly enhancing target detection sensitivity, where the nonlinearly operations are the exponential function with its normalization and the power function with its normalization, respectively. Comparative analysis against seven state-of-the-art hyperspectral target detection methods across four real hyperspectral images demonstrates the effectiveness of the proposed method for hyperspectral target detection, with an increase in detection accuracy and a competitive computational efficiency. An extensive ablation study further validates the critical components of the proposed framework, confirming its comprehensive capability and applicability in hyperspectral target detection scenarios.**

***Index Terms*—Hyperspectral imagery (HSI), momentum contrastive learning, target detection, transformer, unsupervised learning.**

I. Introduction

HYPERSPECTRAL imagery (HSI) is captured by hyperspectral sensors in the visible and short-wave infrared (or mid-wave and long-wave infrared) regions of the spectrum [1], [2], which not only contains the spatial information of the scene but also collects the spectral information of the ground objects to form the image cube data of three dimensions, with two spatial dimensions of the scene, and one spectral dimension consisting of the characteristics of the electromagnetic wave reflection signal at a specific wavelength [3]. The spectrum of each pixel in the HSI can reflect the reflection characteristics of different ground objects in the scene [4]. Benefiting from the high spectral resolution of HSIs [5], hyperspectral target detection (HTD) can detect targets based on the spectral differences of different ground objects and has essential applications in the fields of military camouflage target identification [6], [7], pollution detection [8], [9], mineral exploration [10], food safety [11], and medical diagnosis [12].

HTD has been developed over a long period of time with a large number of classical HTD methods. Spectral matched filtering (SMF) [13] and adaptive coherence estimation (ACE) [14] are classical HTD methods based on probabilistic statistics assuming that the background conforms to a multivariate Gaussian distribution. The constrained energy minimization (CEM) [15] method highlights the target and suppresses the background by designing a finite pulse filter that minimizes the overall energy output under the constraints of the target signal. The orthogonal subspace projection (OSP) [16] method achieves HTD by projecting the target onto the orthogonal subspace of the background subspace and then maximizing the signal-to-noise ratio on the projection subspace. However, these HTD methods based on linear spectral information do not explore the nonlinear relationship between spectral bands. Therefore, kernel-based learning theory is used for HTD to exploit the nonlinear correlations of the HSI data. Some classical HTD methods have been extended to the corresponding kernel-based nonlinear versions, such as kernel SMF [17], kernel ACE [18], kernel CEM [19], kernel OSP [20], etc. In most cases, the kernel-based HTD methods assume that linearly inseparable data in low-dimensional space

Manuscript received 4 January 2024; revised 16 February 2024 and 17 March 2024; accepted 1 April 2024. Date of publication 12 April 2024; date of current version 1 May 2024. This work was supported in part by the National Nature Science Foundation of China under Grant 61801075 and Grant 42271355, in part by the Natural Science Foundation of Liaoning Province under Grant 2022-MS-160, in part by the China Postdoctoral Science Foundation under Grant 2020M670723, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132023238. *(Corresponding author: Xi Chen.)*

Yulei Wang, Xi Chen, Enyu Zhao, Meiping Song, and Chunyan Yu are with the Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: wangyulei@dlmu.edu.cn; xi_chen@nudt.edu.cn; zhaoenyu@dlmu.edu.cn; smping@163.com; yuchunyan1997@126.com).

Chunhui Zhao is with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150009, China (e-mail: zhaochunhui@hrbeu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3387985

will likely become linearly separable in high-dimensional space. Recently, HTD methods based on sparse representation have been proposed successively. Chen et al. [21] first proposed a sparsity-based target detection method (STD), which represents the pixels to be detected by a linear approximation of the atomic vectors in the complete dictionary, then calculates the reconstruction error with the pixels to be detected, and finally determines whether the pixels to be detected are targeted by a set threshold. Li et al. [22] proposed a combined sparse and collaborative representation (CSCR) of the HTD method, which implements target detection by representing the pixels to be detected with a target library and a background library. However, the representation-based HTD methods require prior information to construct the dictionary, which is difficult to obtain in practical applications.

Due to the strong generalization and deep extraction of advanced semantic features, deep learning has been gradually applied in HSI processing [23], [24]. In recent years, deep learning based HTD algorithms have gradually been proposed. For HTD tasks, normally the prior information is only a spectrum of the target of interest, and it is not possible to train the deep neural network in a supervised manner directly based on the prior target spectrum. From the perspective of transfer learning, some methods transfer the model knowledge trained on the dataset with known labels to the target detection task, such as the convolutional neural network-based detection (CNND) [25] method, the spectral-spatial joint target detection method of hyperspectral image based on transfer learning [26], the sensor-independent HTD (SIHTD) method [27], and the meta-learning and Siamese network-based HTD (MLSN) [28] method. CNND pairs and assigns label 0 between similar pixels spectra and label 1 between different classes of pixels spectra based on the known labeled information from a hyperspectral dataset with known labels in the source domain and then trains a binary-classified multilayer CNN for HTD using the samples generated by the pixel pairing. However, the unmatched hyperspectral sensors in the source and target domains [29] can seriously affect the performance of transfer learning on HTD. To solve this problem, SIHTD adaptively transfers the similarity and dissimilarity measurement from the source domain to the target domain for HTD in an adversarial manner. Some methods start from the perspective of expanding the training samples. A deep CNN for HTD (denoted as HTD-Net) [30] uses a modified autoencoder with a contracting path and a symmetric expanding path to generate target signatures, where the background samples significantly different from the target samples are found based on the linear prediction strategy, and then the obtained target and background samples are paired to train the deep CNN to learn the spectral differences between the paired samples. An HTD method with an auxiliary generative adversarial network [31] expands the training set by generating simulated target and background spectra using a generative adversarial network. A two-stream convolutional network-based target detector [32] finds enough typical background pixels by a hybrid sparse representation and classification-based pixel selection strategy, and then pairs the prior target with the synthesized target and background samples, respectively, to form positive and negative sample pairs to train a binary classification network. Rao et al. [33] proposed a Siamese transformer network for HTD, which extracts the high-purity background pixels in the HSI to be detected by endmember extraction and unmixing algorithms. There are also deep learning-based HTD methods that rely on prior information obtained from traditional HTD methods to help model learning. Shi et al. [34] proposed a method for HTD using region of interest (ROI) feature transformation and multiscale spectral attention, where the ROI map is obtained by a CEM detector and an edge-preserving filter, and the HSI to be detected is fed with the ROI map into a constructed deep spatial-spectral network for extracting spatial and spectral features of interest, and then the HTD detection results are obtained using the nearest neighbors. The background learning based target suppression constraint (BLTSC) [35] detector finds reliable background samples for training adversarial autoencoder (AAE) by performing coarse detection by CEM detector, and then reconstructs the original HSI using the well-trained AAE, and finally, the discrepancy between the reconstructed and original HSIs are examined to spot the targets.

In summary, the performance of transfer learning-based HTD methods is primarily limited by the adaptability of the transferred knowledge. The performance of HTD methods that help model training with the help of the prior information obtained from traditional HTD methods can be limited by the performance of traditional HTD methods. The HTD methods that expand the training samples by pairing or mixing some target and background samples found from the HSI to be detected will be affected by the quality of the target and background samples found. The HTD methods using CNNs obtain the approximate global information of the spectrum by building a deep CNN.

In recent years, contrastive learning has been widely applied as an unsupervised representation learning method in various fields, including computer vision. Such as a simple framework for contrastive learning of visual representations (SimCLR) [36], unsupervised learning of visual features by contrasting cluster assignments (SwAV) [37], momentum contrast for unsupervised visual representation learning (MoCo) [38], and a new method BYOL for self-supervised image representation learning without negative sample pairs proposed in [39]. SimCLR directly uses negative samples coexisting in the current batch, and it requires a large batch size to work well, and MoCo maintains a queue of negative samples and turns one branch into a momentum encoder to improve consistency of the queue [40]. Their competitive performance in downstream tasks brings a strong theoretical support for HTD methods oriented to contrastive learning.

To overcome the reliance on explicit target and background samples, this article proposes a novel unsupervised learning framework based on unsupervised momentum contrast learning and transformer (MCLT). Unlike existing approaches, the proposed method innovatively combines transformer-based and momentum encoder networks for enhanced spectral feature extraction, capitalizing on both the detailed local and the global spectral information. Furthermore, the application of the unsupervised momentum contrastive learning, complemented by a strategic queuing mechanism for negative sample management, sets a new standard for feature discriminability in the absence

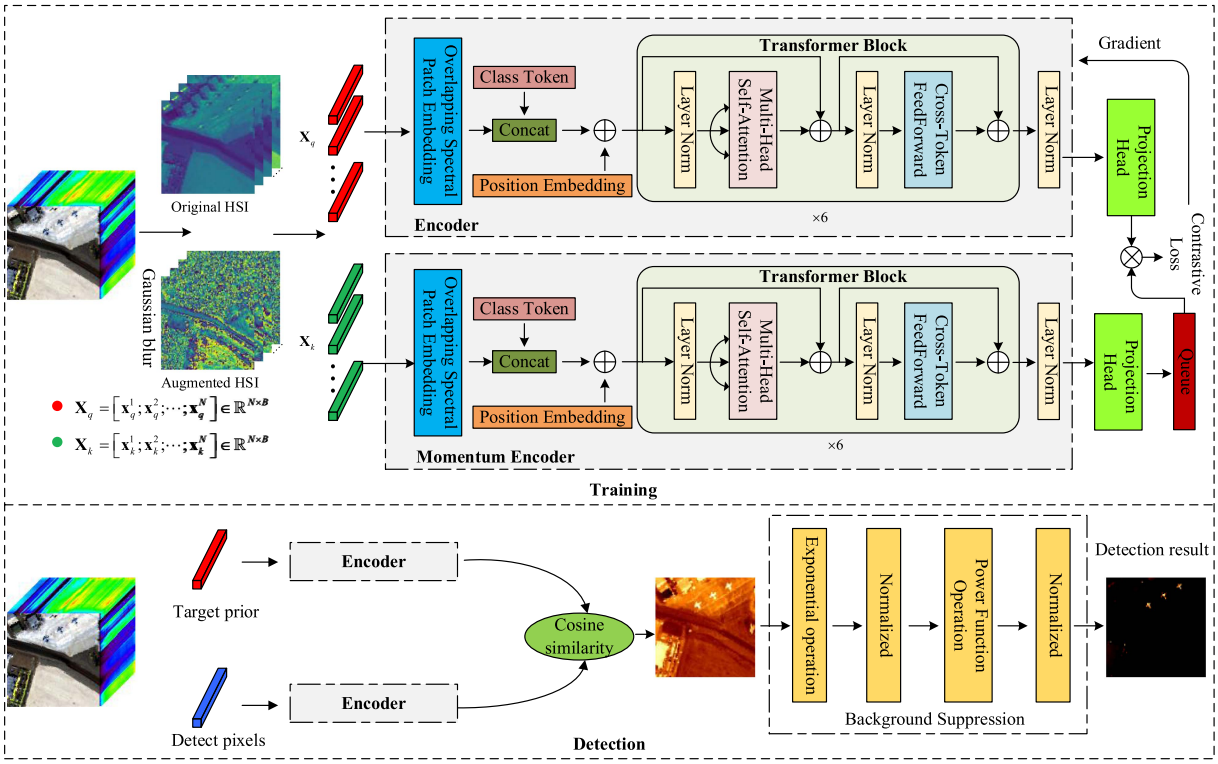

Fig. 1. Overall flowchart of the proposed HTD method based on unsupervised MCLT.

of labeled data. In addition, the proposed background suppression technique, utilizing nonlinear transformations, significantly improves detection sensitivity and accuracy. The main contributions of this article can be summarized as follows, where these contributions not only fill a critical gap in the literature but also surpass existing methods, as evidenced by the comprehensive comparative analysis and ablation study.

1) For spectral target detection, a novel encoder design that integrates transformer-based and momentum encoding to capture both local and global spectral features, addressing the oversight of local spectral detail in existing models.
2) Unsupervised momentum contrastive learning equips the model with the ability to discriminate differences between spectra, freeing it from dependence on labeled target and background samples.
3) An innovative background suppression technique that leverages nonlinear transformations is proposed for a better separation of background and target pixels, where exponential and normalization operations, and power function and normalization operations are used.

The rest of this article is organized as follows. Section II gives a detailed description of the proposed MCLT method. The experimental studies and analysis to verify the proposed method are presented in Section III. Finally, the conclusions are drawn in Section IV.

II. Proposed Method

This section delineates the proposed MCLT method in a comprehensive manner. Fig. 1 presents the methodological flowchart of the MCLT approach, encapsulating its systematic workflow. The methodology primarily unfolds in three sequential steps: the construction of a transformer-based encoder tailored for HTD, the implementation of spectral discriminability learning with momentum encoder, and the execution of background suppression. Notably, the initial two steps are encompassed within the training phase, aimed at preparing the model by enhancing its ability to distinguish between spectral signatures. The final step is situated within the detection phase, strategically designed to optimize target-background separation. This structured approach underscores the MCLT method's capability to effectively identify and isolate targets from complex hyperspectral backgrounds.

A. Transformer-Based Encoder for HTD

The encoder module in Fig. 1 depicts the architecture of the transformer-based encoder designed specifically for HTD. This encoder structure is pivotal in extracting and processing the rich spectral and spatial information inherent in hyperspectral images for effective target detection. The architecture comprises three main components: overlapping spectral patch embedding, position embedding, and the transformer block, each of which plays a critical role in the encoder's functionality, with detailed processing given in the following.

1) Overlapping Spectral Patch Embedding and Position Embedding: Transformer was first designed for machine translation tasks [41], using self-attention mechanisms to process sequence data. Since then it has been widely used in natural language processing (NLP), such as BERT [42]. Due to the success of

using transformer in the field of NLP, transformer has been concerned to be applied in the field of computer vision in recent years, such as Vision Transformer (ViT) [43], Swin Transformer [44], et al. ViT divides the input image into nonoverlapping image blocks and linearly projects each image block into a *d*-dimensional feature vector using the learnable weight matrix [45]. Inspired by ViT, the spectrum is divided into several patches of the same sequence length as the input of transformer to reduce the length of the input sequence, facilitating straightforward processing and analysis with lower computational complexity. However, such nonoverlapping spectral patches would overlook local information between adjacent spectral patches when performing self-attention operations, potentially leading to information loss or suboptimal performance. Therefore, overlapping spectral patch embedding is designed to provide higher quality token sequences to improve the performance of transformer.

The overlapping spectral patch embedding divides the spectrum into a number of spectral patches of fixed sequence length with overlapping regions, capturing spectral relationships between neighboring patches, enabling the transformer to extract the global information of the spectrum while focusing on the local detail information of the spectrum. It is implemented using a one-dimensional (1-D) convolutional layer with the number of convolutional kernels *d*, a convolutional kernel size *k*, a step size *s*, and no zero padding for overlapping spectral patch segmentation and feature mapping. The number of convolution kernels *d* controls the dimensionality of each spectral patch after feature mapping, the size of the convolution kernel *k* controls the length of each spectral patch, and the step size *s* controls the size of the nonoverlapping part between adjacent spectral patches.

For a pixel spectrum $\mathbf{x}$ with band *B*, the embedded spectral token sequence $\mathbf{x}_e \in \mathbb{R}^{N\times d}$ is obtained by overlapping spectral patch embedding, where $N = ((B\text{-}k)/s{+}1)$ is the effective input sequence length of the transformer and *d* is the dimension of each embedded spectral token sequence. Then, a learnable embedding $\mathbf{x}_{\text{learn}}$ is added before the embedded spectral token sequence $\mathbf{x}_e$, and the output of $\mathbf{x}_{\text{learn}}$ obtained by the transformer block is used as the representation of the spectrum of this pixel. Finally, the learnable 1-D position embeddings are added to the embedded spectral token sequence to retain the position information of the spectral patch in the original pixel spectrum. The final embedded spectral token sequence as the transformer block input can be expressed as follows:

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{learn}}; \mathbf{x}_e^1; \mathbf{x}_e^2; \cdots; \mathbf{x}_e^N\right] + \mathbf{E}_{pos} \tag{1}$$

where $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times d}$ is the learnable 1-D position embedding.

2) Transformer Block: The transformer block, depicted in the transformer block segment of Fig. 1, embodies a structured framework comprising essential components including layer normalization (LN) [46], multihead self-attention (MSA) [41], residual connection [47], and cross-token feedforward layer (CTFFL).

For a given input sequence, denoted as $\mathbf{z} \in \mathbb{R}^{N\times d}$, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are derived through the projection of a learnable feature matrix. This process can be succinctly delineated as follows:

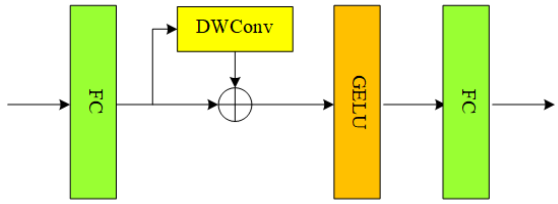

Fig. 2. Cross-token feedforward layer.

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{z}\mathbf{E}_{QKV} \tag{2}$$

where $\mathbf{E}_{QKV} \in \mathbb{R}^{d\times 3d_k}$ signifies the learnable feature matrix.

The dot product between $\mathbf{Q}$ and $\mathbf{K}$ is computed and subsequently scaled through divided by $\sqrt{d_k}$ to mitigate potential gradient issues arising from large dot product values. The weight of $\mathbf{V}$ is determined via the softmax function, which is multiplied by value $\mathbf{V}$, yielding the self-attention mechanism for the embedded spectral token sequence, formulated as follows:

$$\text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where d_k represents the dimension of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$.

The MSA mechanism operates by conducting *h* self-attention operations, called "heads," in parallel, followed by projecting their concatenated outputs. The MSA can be formalized as follows:

$$[\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i] = \mathbf{z}\mathbf{E}_{Q_iK_iV_i}, i = 1 \ldots h \tag{4}$$

$$\begin{aligned}\text{MSA}(\mathbf{z}) = [&\text{SA}_1(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1); \text{SA}_2(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2); \cdots \\ &; \text{SA}_h(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h)]\,\mathbf{E}_{MSA}\end{aligned} \tag{5}$$

where $\mathbf{E}_{Q_iK_iV_i} \in \mathbb{R}^{d\times 3d_k}$ denotes the learnable projection matrix employed to project the input sequence into queries $\mathbf{Q}_i$, keys $\mathbf{K}_i$, and values $\mathbf{V}_i$, $\mathbf{E}_{MSA} \in \mathbb{R}^{h\cdot d_k\times d}$ represents the feature matrix of the projection concatenated output, and *h* is the number of self-attention operations used in parallel. To make the total computational cost of MSA similar to that of full-dimensional single-head self-attention, d_k is generally set to *d*/*h*. The transformer block is repeated *L* times to build an encoder with *L* layers. The output of MSA in each transformer block layer can be formalized as follows:

$$\mathbf{z}_l' = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1},\ l = 1 \ldots L. \tag{6}$$

The feedforward layer within the original transformer encoder is a fully connected feedforward network consisting of two linear transforms with a ReLU activation function in [48]. Its fully connected layer is point-wise and cannot learn cross-token information [49]. The CTFFL is designed to complement the local detail information in the feedforward layer, as shown in Fig. 2.

CTFFL enhances the capture of local details within the feedforward layer by incorporating depth-wise convolution between the two fully connected layers of the feedforward layer. The

process can be described as follows:

$$\mathbf{Z}' = \mathrm{FC}\left(\mathbf{Z};\omega_1\right) \tag{7}$$

$$\mathbf{Z}'' = \mathrm{FC}\left(\sigma\left(\mathbf{Z}' + \mathrm{DWConv}\left(\mathbf{Z}';\omega\right)\right);\omega_2\right) \tag{8}$$

where $\mathbf{Z}$ is the input of the CTFFL, ω_1 and ω_2 are the parameters of the two fully connected layers, ω is the parameter of the 1-D depth-wise convolutional layer, and σ is the Gaussian error linear unit [50] activation function. The output of the CTFFL in each transformer block layer can be formalized as follows:

$$\mathbf{z}_l = \mathrm{CTFFL}\left(\mathrm{LN}\left(\mathbf{z}_l'\right)\right) + \mathbf{z}_l', l = 1 \ldots L. \tag{9}$$

This work uses $h = 8$ parallel self-attention operations and $L = 2$ layers of transformer block. The output of the encoder can be expressed as follows:

$$\mathbf{y} = \mathrm{LN}\left(\mathbf{z}_L^0\right) \tag{10}$$

where $\mathbf{z}_L^0$ is the output of the corresponding position obtained after the learnable embedding $\mathbf{x}_{\text{learn}}$ passes through the transformer block, and the output $\mathbf{y}$ after $\mathbf{z}_L^0$ passes through the LN is used as the representation of the spectrum.

B. Spectral Discriminability Learning

Spectral discriminability learning constructs pretext tasks for spectral instance discrimination by data augmentation and trains the model with unsupervised momentum contrastive learning [51] to obtain a discriminative encoder with spectral difference discrimination for HTD.

Data augmentation is achieved by applying Gaussian blur [52] to the original HSI $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$. Gaussian blurring is achieved by convolving each band in the original HSI with a Gaussian kernel. We randomly sampled the standard deviation $\delta = [0.1, 2.0]$, with the kernel size set to 1×1. The spectral instance discrimination pretext task is achieved by pairing the original HSI with the spectra of pixels at the same location in the HSI after performing Gaussian blurring. The spectra of pixels at the same position in the original HSI and the HSI after performing data augmentation can be considered as positive pairs. The spectral instance discrimination pretext task is constructed to generate a self-supervised signal to help train a discriminative encoder.

The flow of spectral discriminability learning is shown in the training part of Fig. 1. A mini-batch of pixel spectra $\mathbf{X}_q$ is randomly sampled from the original HSI, and the augmented samples $\mathbf{X}_k$ are obtained after data augmentation, expressed as $\mathbf{X}_q = [\mathbf{x}_q^1; \mathbf{x}_q^2; \cdots; \mathbf{x}_q^N] \in \mathbb{R}^{N\times B}$, $\mathbf{X}_k = [\mathbf{x}_k^1; \mathbf{x}_k^2; \cdots; \mathbf{x}_k^N] \in \mathbb{R}^{N\times B}$.

The representation of $\mathbf{X}_q$ is extracted using the encoder $f_{\text{encoder}}(\cdot)$, and the feature matrix $\mathbf{U}_q = [\mathbf{u}_q{}^1, \mathbf{u}_q{}^2, \ldots, \mathbf{u}_q{}^N] \in \mathbb{R}^{N\times d_{\text{MLP}}}$ is obtained after mapping through the projection head. It can be formalized as follows:

$$\mathbf{U}_q = \mathrm{MLP}_{\text{encoder}}\left(f_{\text{encoder}}\left(\mathbf{X}_q;\theta_{\text{encoder}}\right);\theta_{\mathrm{MLP}_{\text{encoder}}}\right). \tag{11}$$

The representation of $\mathbf{X}_k$ is extracted using the momentum encoder $f_{\text{m\_encoder}}(\cdot)$, and the feature matrix is obtained after mapping through the projection head, marked as $\mathbf{V}_k = [\mathbf{v}_k{}^1, \mathbf{v}_k{}^2, \ldots, \mathbf{v}_k{}^N] \in \mathbb{R}^{N\times d_{\text{MLP}}}$. The process can be expressed as follows:

$$\mathbf{V}_k = \mathrm{MLP}_{\text{m\_encoder}}\left(f_{\text{m\_encoder}}\left(\mathbf{X}_k;\theta_{\text{m\_encoder}}\right);\theta_{\mathrm{MLP}_{\text{m\_encoder}}}\right) \tag{12}$$

where θ_{encoder} and $\theta_{\text{m\_encoder}}$ are the parameters of the encoder $f_{\text{encoder}}(\cdot)$ and the momentum encoder $f_{\text{m\_encoder}}(\cdot)$, respectively. The projection head $\mathrm{MLP}_{\text{encoder}}$ of the encoder and the projection head $\mathrm{MLP}_{\text{m\_encoder}}$ of the momentum encoder are MLP containing a hidden layer with parameters $\theta_{\mathrm{MLP}_{\text{encoder}}}$ and $\theta_{\mathrm{MLP}_{\text{m\_encoder}}}$, respectively. $\mathrm{MLP}_{\text{encoder}}$ and $\mathrm{MLP}_{\text{m\_encoder}}$ use the ReLU activation function.

Then the feature $\mathbf{V}_k = [\mathbf{v}_k{}^1, \mathbf{v}_k{}^2, \ldots, \mathbf{v}_k{}^N] \in \mathbb{R}^{N\times d_{\text{MLP}}}$ outputted by the momentum encoder through the projection head is fed into the queue. The queue has the property of first-in-first-out. The feature samples in the queue are considered as negative samples. The feature samples in the queue are gradually replaced, with the current mini-batch of feature samples entering the queue and the oldest mini-batch of feature samples leaving the queue. The negative sample size used for contrastive learning can be separated from the mini-batch input sample size by the queue. Therefore, the negative samples could be stored by setting a large queue. A large queue can help the model to learn more discriminative features since it contains abundant negative samples. The queue size can be flexibly and independently set to the hyperparameter K.

Finally, the output features of the encoder $f_{\text{encoder}}(\cdot)$ through the projection head $\mathrm{MLP}_{\text{encoder}}$ are fed into the contrastive loss with the features in the queue. The contrastive loss maximizes the similarity of positive pairs and minimizes the similarity of negative pairs, enabling the encoder to have spectral difference discrimination capability by optimizing the contrastive loss. In this article, the similarity between positive and negative pairs is measured by the dot product, and the contrastive loss uses the InfoNCE loss function [53], expressed as follows:

$$L_{\text{InfoNCE}} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{\exp\left(\mathbf{u}_q^i \cdot \mathbf{v}_k^i/\tau\right)}{\sum_{j=0}^{K}\exp\left(\mathbf{u}_q^i \cdot \mathbf{v}_k^j/\tau\right)} \tag{13}$$

where τ is the temperature hyperparameter, $\mathbf{v}_k^j$ includes a positive sample embedding feature (assuming $\mathbf{v}_k^0 = \mathbf{v}_k^j$) and K negative sample embedding features.

During training, in order to keep the features in the queue stay in step, the features in the queue should be generated using the same or similar momentum encoder and projection head, thus to help the model avoid learning to shortcut solutions. Therefore, momentum is used to update the momentum encoder $f_{\text{m\_encoder}}(\cdot)$ and its projection head $\mathrm{MLP}_{\text{m\_encoder}}$. The process can be formalized as follows:

$$\theta_{\text{m\_encoder}} \leftarrow m\theta_{m\_encoder} + (1-m)\,\theta_{\text{encoder}} \tag{14}$$

$$\theta_{\mathrm{MLP}_{\text{m\_encoder}}} \leftarrow m\theta_{\mathrm{MLP}_{\text{m\_encoder}}} + (1-m)\,\theta_{\mathrm{MLP}_{\text{encoder}}} \tag{15}$$

where $m \in [0, 1]$ is the momentum coefficient, set to 0.999 in this article. It is important to note that the encoder $f_{\text{encoder}}(\cdot)$ is updated with its projection head $\mathrm{MLP}_{\text{encoder}}$ by gradient backpropagation.

C. Target Detection and Background Suppression

The procedure of target detection and background suppression is shown in the detection part of Fig. 1.

1) Target Detection: The target prior $\mathbf{x}_t \in \mathbb{R}^{1\times B}$ is compared with each pixel spectrum in the original HSI expressed as $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_{H\times W}] \in \mathbb{R}^{(H\times W)\times B}$ by extracting corresponding representation through a well-trained encoder $f_{\text{encoder}}(\cdot)$. The similarity of the representations between each pixel under test in the original HSI and the target prior is then measured by cosine similarity to obtain the target detection result $\mathbf{B} = [b_1; b_2; \cdots; b_{H\times W}]$, calculated as follows:

$$b_i = \frac{f_{\text{encoder}}(\mathbf{x}_i) \cdot f_{\text{encoder}}(\mathbf{x}_t)^{\mathrm{T}}}{\|f_{\text{encoder}}(\mathbf{x}_i)\| \, \|f_{\text{encoder}}(\mathbf{x}_t)\|}. \tag{16}$$

2) Background Suppression: The values of the target pixels in the detection result $\mathbf{B}$ obtained by cosine similarity are relatively large with significance, however, the distance difference between the values of background and target pixels is relatively small with less significance. The values of the background pixels in $\mathbf{B}$ can be kept away from the values of the target pixels by the exponential and normalization operations. Then the values of the background pixels are further kept away from the values of the target pixels by the power function and normalization operations to achieve the purpose of background suppression. Background suppression is achieved by exponential and normalization operations, power function, and normalization operations, which can be represented as follows:

$$\mathbf{S} = \alpha^{\mathbf{B}} \tag{17}$$

$$s_i = \frac{s_i - s_{\min}}{s_{\max} - s_{\min}} \tag{18}$$

$$\mathbf{R} = \mathbf{S}^{\beta} \tag{19}$$

$$r_i = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{20}$$

where α and β are positive parameters that can adjust the background suppression performance.

III. Experimental Results and Analysis

A. Datasets Description

The experiments are conducted on four real hyperspectral images with different scenarios, where the datasets are obtained by three different HSI sensors.

1) San Diego Dataset: The San Diego dataset was collected by an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) at San Diego airport, CA, USA. It consists of 224 bands with wavelengths ranging from 370 to 2510 nm. Due to low signal-to-noise ratio and water absorption, bands 1–6, 33–35, 97, 107–113, 153–166, and 221–224 were removed, leaving the rest 189 bands for HTD. The whole image has 400 × 400 pixels. The spatial resolution is 3.5 m, and the spectral resolution is 10 nm. In the experiment, two scene regions of size 120 × 120 and 100 × 100, named as San Diego A and San Diego B, are intercepted from the upper left corner and the center of the San Diego dataset, respectively. The plane pixels in San Diego A and San Diego B scenes are considered targets for HTD and contain 58 and 134 target pixels, respectively. The pseudo-color images of San Diego A and San Diego B with ground truth are shown in Figs. 3(a)-(b) and 4(a)-(b), respectively.

2) PaviaC Dataset: The PaviaC dataset was captured by the Reflection Optical System Imaging Spectrometer (ROSIS-03) in the central city of Pavia, Italy. It has 100 × 120 pixels and contains 102 bands with wavelengths ranging from 430 to 860 nm. The spatial resolution is 1.3 m, and the spectral resolution is 4 nm. The background in this scene is mainly composed of water and bridge, and the vehicles on the bridge are considered as the targets for HTD with a total of 68 pixels. Fig. 5(a) and (b) show the pseudo-color image and ground truth of the PaviaC dataset.

3) MUUFL Gulfport Dataset: The MUUFL Gulfport dataset [54], [55] was collected in November 2010 at the University of Southern Mississippi Gulf Park campus in Long Beach, Mississippi. The size of the original dataset is 325 × 337 pixels with 72 bands. The first four and last four bands were removed due to noise, yielding a new HSI with 64 bands. The lower right corner of the original HSI contains invalid regions, so only the first 220 columns are used for ground truth mapping. The cropped HSI size was 325 × 220 × 64, with a total of 269 clothing panel pixels in the scene considered as target for HTD. Fig. 6(a) and (b) show the pseudo-color image and ground truth of the MUUFL Gulfport dataset.

B. Experimental Setup

1) Comparison Methods: A total of seven state-of-the-art HTD methods were used in the experiment to compare the performance with the proposed MCLT. The seven compared methods include two classical HTD methods, CEM [15] and OSP [16], two representation-based methods, CSCR [22] and DM-BDL [56], three deep learning-based methods, BLTSC [35], MLSN [28], and ULMMDL [57]. The comparison methods and the proposed MCLT use the same target prior of the same HSI datasets for HTD.

2) Implementation Details: The proposed MCLT method is implemented by building an encoder for HTD based on transformer, spectral discriminability learning, and background suppression. For the encoder used to extract the spectral representation of each pixel in the HSI to be detected, the parameters (d, k, s) in the overlapping spectral patch embedding are set to (128, 9, 2), (128, 9, 2), (128, 6, 2), and (128, 4, 2) for the San Diego A, San Diego B, PaviaC, and MUUFL Gulfport HSI datasets, respectively. The dimensionality of the embedded spectral token sequence obtained after overlapping spectral patch embedding is $d = 128$. The dimensionality of both the learnable embedding $\mathbf{x}_{\text{learn}}$ and the learnable 1-D location embedding is set to 128. Two transformer blocks are used to construct an encoder with depth of 2 for extracting the representation of each pixel spectrum. MSA in each transformer block is achieved by parallelly running $h = 8$ self-attention operations (called "heads") and projecting their concatenated outputs. The representation obtained by the two-layer transformer block is used as the representation of the entire pixel spectrum. Note that

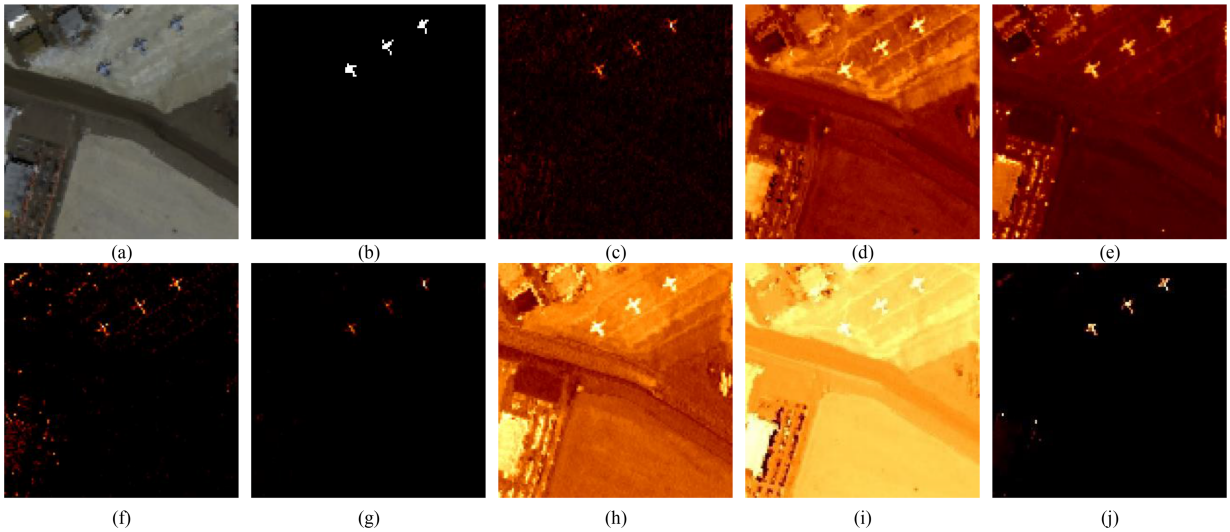

Fig. 3. Detection maps of different methods for San Diego A. (a) Pseudo-color image. (b) Ground truth. (c) CEM. (d) OSP. (e) CSCR. (f) DM-BDL. (g) BLTSC. (h) MLSN. (i) ULMMDL. (j) MCLT.

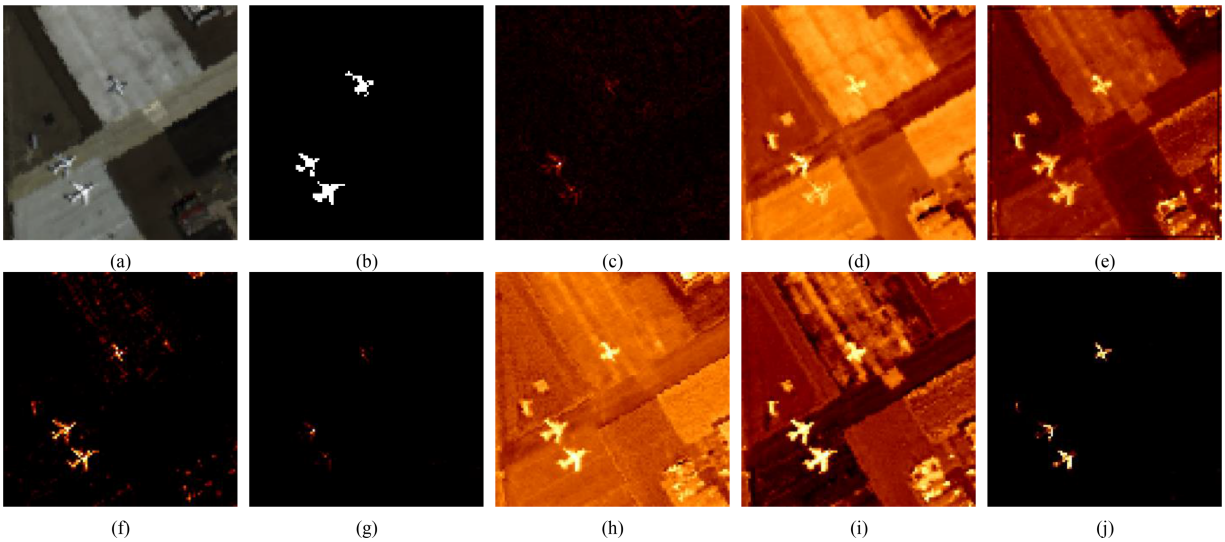

Fig. 4. Detection maps of different methods for San Diego B. (a) Pseudo-color image. (b) Ground truth. (c) CEM. (d) OSP. (e) CSCR. (f) DM-BDL. (g) BLTSC. (h) MLSN. (i) ULMMDL. (j) MCLT.

the input and output of the transformer block have the same dimension. In spectral discriminability learning, the outputs of the encoder and the momentum encoder go through the corresponding projection heads to obtain the features of pixel spectra and their augmented samples. The projection heads of the encoder and the momentum encoder have the same structure, both being a two-layer MLP. The number of neurons in the first and second layers of $\mathrm{MLP}_{\mathrm{encoder}}$ and $\mathrm{MLP}_{\mathrm{m\_encoder}}$ is 128. The queue stores the output of the features by the momentum encoder through the projection head. The queue sizes for San Diego A, San Diego B, PaviaC, and MUUFL Gulfport HSI datasets are set to 7200, 10000, 12000, and 13000, respectively. During training, the learning rate and the temperature coefficient in the contrastive loss are set to 0.5 and 0.07, respectively. The epoch for training is set to 50. The mini-batch during training is set to 480, 400, 600, and 1300 for San Diego A, San Diego B, PaviaC, and MUUFL Gulfport datasets, respectively. For the background suppression process, the exponential operation sets α to 9×10^{47} for all datasets in the experiment, and β in the power function operation is set to 20, 60, 20, and 60 for the San Diego A, San Diego B, PaviaC, and MUUFL Gulfport datasets, respectively. The comparison method follows the settings recommended in the original literature.

The experimental hardware environment consists of an AMD Ryzen Threadripper 3990X 64-core processor with a Quadro RTX 8000 GPU with 48 GB of RAM. Two classical HTD (CEM and OSP) methods and two representation-based HTD (CSCR and DM-BDL) methods are implemented in MATLAB R2017b, and three deep learning-based comparison methods (BLTSC, MLSN, and ULMMDL) are implemented using Python 3.6 and

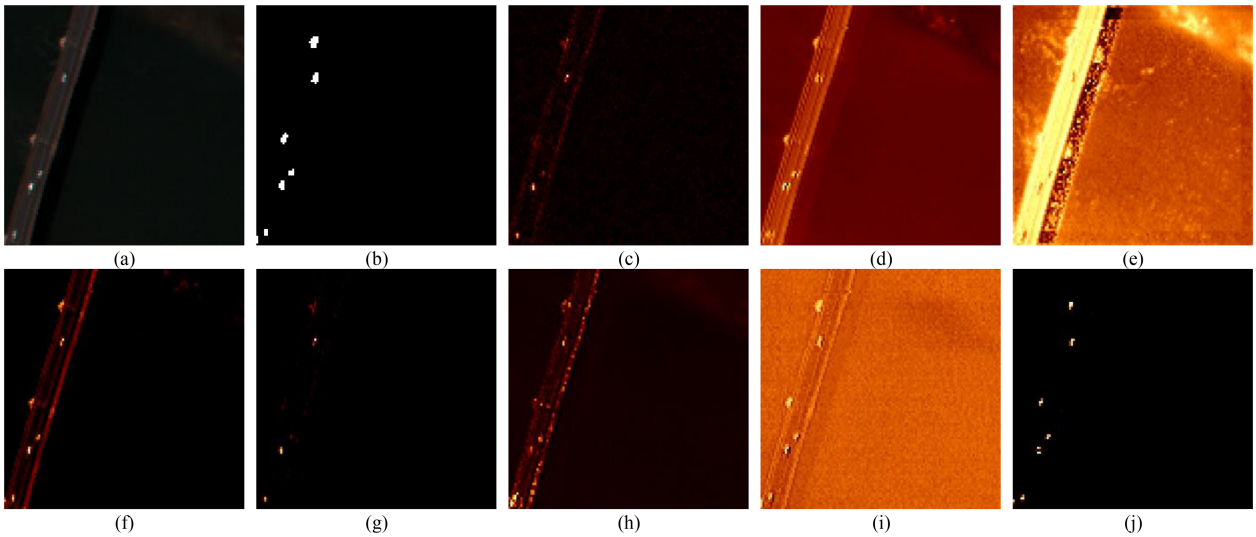

Fig. 5. Detection maps of different methods for PaviaC. (a) Pseudo-color image. (b) Ground truth. (c) CEM. (d) OSP. (e) CSCR. (f) DM-BDL. (g) BLTSC. (h) MLSN. (i) ULMMDL. (j) MCLT.

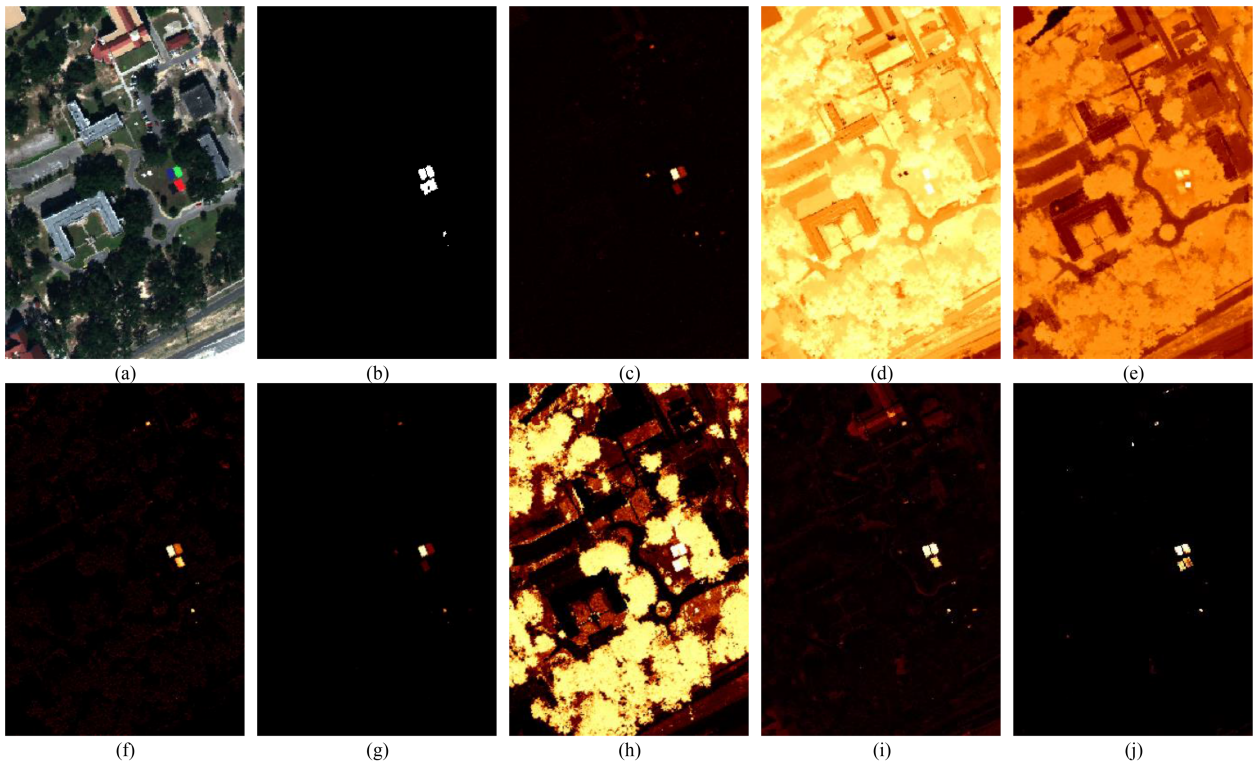

Fig. 6. Detection maps of different methods for MUUFL Gulfport. (a) Pseudo-color image. (b) Ground truth. (c) CEM. (d) OSP. (e) CSCR. (f) DM-BDL. (g) BLTSC. (h) MLSN. (i) ULMMDL. (j) MCLT.

TensorFlow 1.80. The proposed MCLT is implemented using Python 3.8.3 and PyTorch 1.60.

3) Prior Target Spectrum Selection: The prior target spectrum in the experiment was obtained from hyperspectral images, and the coordinates of the prior target spectrum were taken as (11, 88), (36, 49), (57, 14), and (153, 159) for the four datasets in the experiment, respectively. Notably, in each dataset, the prior target spectrum was chosen at the center point of the target, aiming to capture the original spectral properties of the target effectively. Only one target spectrum was taken as the prior target spectrum.

4) Evaluation Criterion: To evaluate the HTD performance of the proposed MCLT method and other comparison methods, 3-D receiver operating characteristic (3-D ROC) curve is used

to measure the performance of the detector. The 3-D ROC curve can be considered as a function of detection probability P_D, false alarm probability P_F, and threshold τ, and can be obtained as the value of τ varies [58]. P_D and P_F can be calculated as follows:

$$P_D(\tau) = \frac{N_{TP,\tau}}{N_{TP,\tau} + N_{FN,\tau}} \tag{21}$$

$$P_F(\tau) = \frac{N_{FP,\tau}}{N_{FP,\tau} + N_{TN,\tau}} \tag{22}$$

where $N_{TP,\tau}$ denotes the number of pixels that are correctly detected as targets at a given threshold τ, $N_{FN,\tau}$ represents the number of pixels that incorrectly detect targets as backgrounds at a given threshold τ, $N_{FP,\tau}$ denotes the number of pixels that incorrectly detect backgrounds as targets at a given threshold τ, and $N_{TN,\tau}$ denotes the number of pixels that are correctly detected as backgrounds at a given threshold τ. Three 2-D ROC curves can be obtained from the 3-D ROC curves, including the 2-D ROC curve of (P_D, P_F), the 2-D ROC curve of (P_D, τ), and the 2-D ROC curve of (P_F, τ), respectively. The 2-D ROC curve of (P_D, P_F) can evaluate the effectiveness of the detector, the 2-D ROC curve of (P_D, τ) can evaluate the target detectability of the detector, and the 2-D ROC curve of (P_F, τ) can evaluate the background suppression ability of the detector. The detector should have better performance with the following three conditions of the corresponding 2-D ROC curves: the closer the 2-D ROC curve of (P_D, P_F) is to the upper left corner of the coordinate axis, the closer the 2-D ROC curve of (P_D, τ) is to the upper right corner of the coordinate axis, and the closer the 2-D ROC curve of (P_F, τ) is to the lower left corner of the coordinate axis.

For quantitative analysis of the detector, the areas under the curves (AUC) of the 2-D ROC curves (P_D, P_F), (P_D, τ), and (P_F, τ) were used as quantitative indicators to quantify the performance of the detector. The detector performs better when AUC(P_D, P_F) and AUC(P_D, τ) are close to 1 and when AUC(P_F, τ) is close to 0. The detection performance improves with higher values of AUC(P_D, P_F) and AUC(P_D, τ), while background suppression improves with lower values of AUC(P_F, τ). By considering these three AUC values, Chang [59] devised a detection measure AUC_{OD} and a background suppression capability measure AUC_{BS} to evaluate the performance of the detector, which was defined as follows:

$$AUC_{OD} = AUC(P_D, P_F) + AUC(P_D, \tau) - AUC(P_F, \tau) \tag{23}$$

$$AUC_{BS} = AUC(P_D, P_F) - AUC(P_F, \tau) \tag{24}$$

where $AUC_{OD} \in [-1, 2]$ and $AUC_{BS} \in [-1, 1]$. The larger the calculated values of AUC_{OD} and AUC_{BS}, the better the detection performance and background suppression effect of the detector.

C. Results and Discussion

Figures (c)–(j) in Figs. 3–6 show the detection maps of the proposed MCLT and other comparison methods for San Diego A, San Diego B, PaviaC, and MUUFL Gulfport datasets, respectively. Subjectively visual assessment from the detection maps, CEM, BLTSC, and the proposed MCLT have good background suppression effect compared with other comparison methods. However, many target pixels are missed in the detection maps of CEM and BLTSC algorithms, only remaining the proposed MCLT with good effect. OSP, CSCR, MLSN, and ULMMDL can detect most of the targets, but the background suppression is not good, making it very difficult to identify them visually. The detection maps of the proposed MCLT method visually show excellent detection performance, with targets highlighted clearly and background suppressed well.

Figs. 7–10 show the 3-D ROC and the corresponding three 2-D ROC curves for the proposed MCLT and seven state-of-the-art comparison methods on the San Diego A, San Diego B, PaviaC, and MUUFL Gulfport datasets corresponding to their detection results. For the 2-D ROC curves of (P_D, P_F) was used to evaluate the detector effectiveness, as shown in Figs. 7(b)–10(b), the 2-D ROC curves of (P_D, P_F) of the MCLT for all HSIs in the experiment are closer to the upper left corner than the comparison methods. For the 2-D ROC curves of (P_D, τ) evaluating the detectability of the detector to the target, as shown in Figs. 7(c)–10(c), the proposed MCLT outperforms CEM and BLTSC, but OSP, CSCR, and ULMMDL perform better than the proposed MCLT. However, for the 2-D ROC curves of (P_F, τ) evaluating the detector background suppression ability, MCLT is very close to the lower left corner and has significantly better background suppression than OSP, CSCR, and ULMMDL.

Since very close ROC curves cannot visually distinguish precisely which detector performs better, the areas under the curves AUC(P_D, P_F), AUC(P_D, τ), and AUC(P_F, τ) of the 2-D ROC curves of (P_D, P_F), (P_D, τ) and (P_F, τ) are used to evaluate the performance of the detectors quantitatively. In addition, AUC_{BS} and AUC_{OD} are used to quantitatively evaluate the background suppression ability and the comprehensive detection performance of the detector. Table I provides specific values of the five AUC measures for MCLT and all comparison methods on the four HSI datasets. The best results in each AUC measure are shown in bold, and the suboptimal results are underlined. Table I shows that the proposed MCLT always obtains the highest AUC(P_D, P_F) of all HSI datasets, verifying its effectiveness in HTD. ULMMDL obtained the highest AUC(P_D, τ) on the San Diego A and San Diego B datasets, demonstrating excellent target detection capabilities. This is made possible by the hierarchical denoising autoencoder (HDAE) designed in the ULMMDL method. HDAE enhances the spectral coherence by iterating over the denoising autoencoder layer by layer, which alleviates the intraclass differences in the target spectra in the HSIs to be detected and makes ULMMDL have a good target preservation capability. However, the values of AUC(P_D, P_F), AUC(P_F, τ), and AUC_{BS} of ULMMDL on four HSI datasets are lower than those of the proposed MCLT method. For the AUC(P_F, τ), BLTSC achieved optimal results on the San Diego A, San Diego B, and MUUFL Gulfport datasets, and MCLT achieved suboptimal results, only slightly weaker than BLTSC. MCLT obtained the optimal results on the PaviaC dataset, and BLTSC got the suboptimal results. BLTSC obtains the weight map of distinguishable targets by background learning. Then the weight map of distinguishable targets is used to correct the

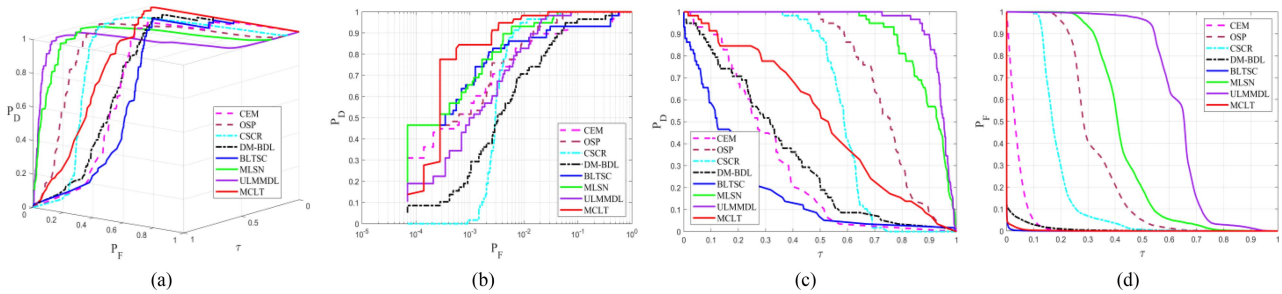

Fig. 7. 3-D ROC and the corresponding 2-D ROC curves of different methods for the San Diego A dataset. (a) 3-D ROC curve. (b) 2-D ROC curve of (P_D, P_F). (c) 2-D ROC curve of (P_D, τ). (d) 2-D ROC curve of (P_F, τ).

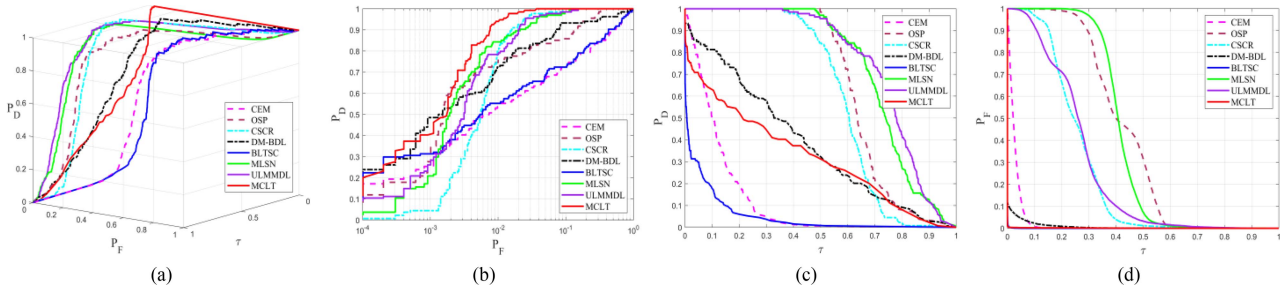

Fig. 8. 3-D ROC and the corresponding 2-D ROC curves of different methods for the San Diego B dataset. (a) 3-D ROC curve. (b) 2-D ROC curve of (P_D, P_F). (c) 2-D ROC curve of (P_D, τ). (d) 2-D ROC curve of (P_F, τ).

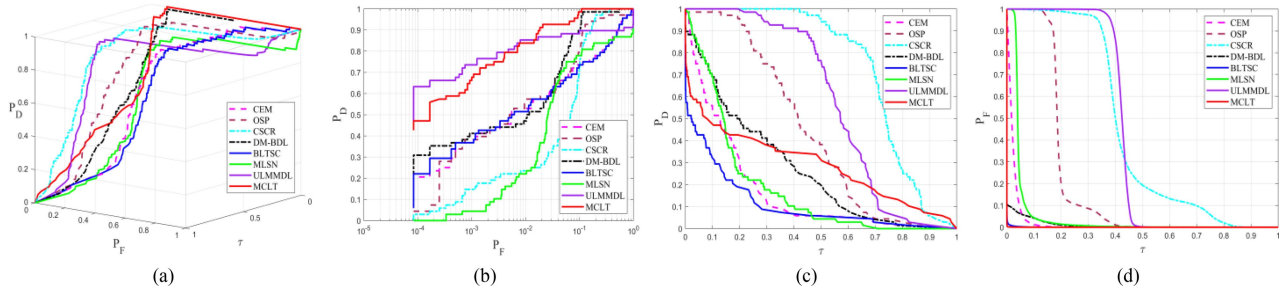

Fig. 9. 3-D ROC and the corresponding 2-D ROC curves of different methods for the PaviaC dataset. (a) 3-D ROC curve. (b) 2-D ROC curve of (P_D, P_F). (c) 2-D ROC curve of (P_D, τ). (d) 2-D ROC curve of (P_F, τ).

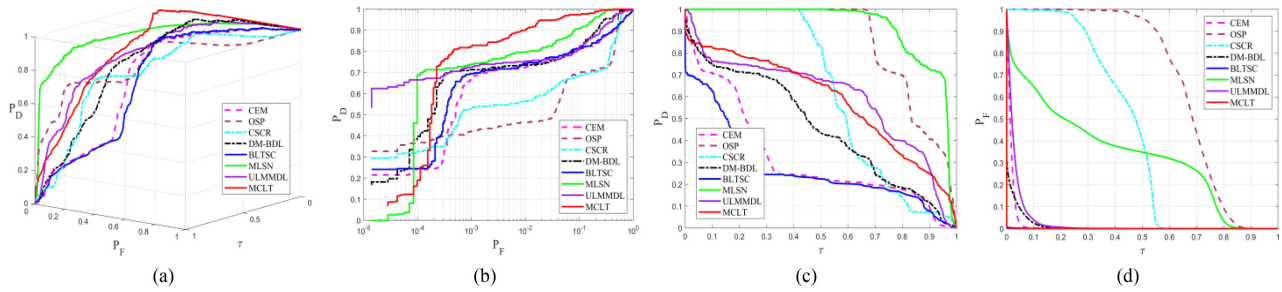

Fig. 10. 3-D ROC and the corresponding 2-D ROC curves of different methods for the MUUFL Gulfport dataset. (a) 3-D ROC curve. (b) 2-D ROC curve of (P_D, P_F). (c) 2-D ROC curve of (P_D, τ). (d) 2-D ROC curve of (P_F, τ).

TABLE I
ACCURACY COMPARISON OF DIFFERENT METHODS FOR FOUR HSI DATASETS

HSI Data Sets		Methods							
		CEM [15]	OSP [16]	CSCR [22]	DM-BDL [49]	BLTSC [33]	MLSN [26]	ULMMDL [50]	MCLT
San Diego A	$AUC_{(P_D,P_F)}$	0.96287	0.99475	0.99546	0.97586	0.96687	0.99688	0.99455	**0.99877**
	$AUC_{(P_D,\tau)}$	0.29733	0.74021	0.58313	0.33058	0.19136	0.89671	**0.94333**	0.51104
	$AUC_{(P_F,\tau)}$	0.03854	0.32051	0.18891	0.00874	**0.00063**	0.42555	0.63472	0.00192
	AUC_{BS}	0.92433	0.67424	0.80655	0.96712	0.96624	0.57133	0.35983	**0.99685**
	AUC_{OD}	1.22166	1.41445	1.38968	1.29770	1.15760	1.46804	1.30316	**1.50789**
San Diego B	$AUC_{(P_D,P_F)}$	0.87253	0.96649	0.99151	0.95445	0.88019	0.99164	0.99143	**0.99704**
	$AUC_{(P_D,\tau)}$	0.12237	0.65370	0.60090	0.37670	0.05052	0.73711	**0.75871**	0.31875
	$AUC_{(P_F,\tau)}$	0.02637	0.41919	0.25346	0.00739	**0.00019**	0.41285	0.25950	0.00082
	AUC_{BS}	0.84616	0.54730	0.73805	0.94706	0.88000	0.57879	0.73193	**0.99622**
	AUC_{OD}	0.96853	1.2010	1.33895	1.32376	0.93052	1.31590	**1.49064**	1.31497
PaviaC	$AUC_{(P_D,P_F)}$	0.87176	0.93250	0.92624	0.96398	0.86545	0.83108	0.89930	**0.99126**
	$AUC_{(P_D,\tau)}$	0.16579	0.43401	**0.73520**	0.26354	0.10956	0.17797	0.57599	0.28855
	$AUC_{(P_F,\tau)}$	0.02262	0.20101	0.44477	0.00956	0.00051	0.04926	0.42257	**0.00007**
	AUC_{BS}	0.84914	0.73149	0.48147	0.95442	0.86494	0.78182	0.47673	**0.99119**
	AUC_{OD}	1.01493	1.16550	1.21667	1.21796	0.97450	0.95979	1.05272	**1.27974**
MUUFL Gulfport	$AUC_{(P_D,P_F)}$	0.90268	0.85376	0.85760	0.94445	0.90693	0.97543	0.93750	**0.99171**
	$AUC_{(P_D,\tau)}$	0.31919	0.85032	0.62237	0.45658	0.27066	**0.91938**	0.59534	0.58308
	$AUC_{(P_F,\tau)}$	0.01624	0.67801	0.43781	0.01187	**0.00015**	0.33405	0.02928	0.00050
	AUC_{BS}	0.88644	0.17575	0.41979	0.93258	0.90678	0.64138	0.90822	**0.99121**
	AUC_{OD}	1.20563	1.02607	1.04216	1.38916	1.17744	1.56076	1.50356	**1.57429**

Bold-font highlights the best result, while underline the second.

results of CEM coarse detection to detect targets and suppress background. However, the performance of BLTSC relies on the performance of the coarse detection method. MCLT does not rely on the prior information found by traditional methods, and the $AUC(P_D, P_F)$, $AUC(P_D, \tau)$, AUC_{BS}, and AUC_{OD} of MCLT are better than those of BLTSC on the four HSIs in the experiment. For the AUC_{BS} used to combine the effects of P_D and P_F on background suppression, the proposed MCLT achieved optimal results on four HSI datasets. The excellent background suppression shows that two nonlinear pull-ups by exponential and power function operations can effectively suppress the background and preserve the target. For the AUC_{OD} used to evaluate the overall detection performance of the detector, the MCLT achieved the best overall performance on San Diego A, PaviaC, and MUUFL Gulfport. This shows that MCLT achieves competitive results with unsupervised momentum contrastive learning for spectral discriminability learning and an encoder based on transformer constructed for extracting spectral features of pixels.

Fig. 11 shows the target-background separability box plots of the detection results of the MCLT and comparison methods on the four HSI datasets. In the target-background separability box plots, target and background pixels with statistically distributed values are placed in the box, removing the highest and lowest 10% of data in the target and background classes [60]. The red boxes indicate the distribution of targets, and the green boxes indicate the distribution of backgrounds. In the boxes, the middle horizontal line indicates the median value. The horizontal lines at the top and bottom rows of each box indicate the maximum and minimum values. The target-background separability box plot not only reflects the separability of the target and background in the detection results but also observes the distribution range of the target and background pixels detection values in the detection results. As can be seen in Fig. 11, for the four HSI datasets, the MCLT suppresses the detected values of the background pixels in the detection results to zero, demonstrating excellent background suppression, and the MCLT also separates the target from the background well. Competitive separability suggests that unsupervised momentum contrastive learning enables the model to learn spectral difference discrimination effectively and enables the model to distinguish well between targets and backgrounds in HSIs to be detected.

D. Ablation Studies

1) Effect of Overlapping Spectral Patch Embedding on Target Detection Accuracy: To investigate the effect of overlapping spectral patch embedding on the HTD accuracy, the overlap between adjacent spectral patches is changed to observe the impact of different sizes of overlapping patches on the HTD accuracy. For the San Diego A and San Diego B datasets with 189 bands, each pixel spectrum is divided into spectral patches of length 9. For the PaviaC dataset with 102 bands, each pixel spectrum is divided into spectral patches of length 6. For the MUUFL Gulfport dataset with 64 bands, each pixel spectrum is divided into spectral patches of length 4. Fig. 12(a)–(d) show the effect of different size overlaps on HTD accuracy versus time consumption on the San Diego A, San Diego B, PaviaC, and MUUFL Gulfport datasets, respectively. The horizontal

Fig. 11. Target-background separability box plots on four HSIs. (a) San Diego A. (b) San Diego B. (c) PaviaC. (d) MUUFL Gulfport.

coordinates of each subfigure in Fig. 12 indicate the length of the divided spectral patches and the length of the nonoverlapping part between adjacent spectral patches. The overlap between adjacent spectral patches in the horizontal coordinate decreases from left to right until there is no overlap between adjacent spectral patches. For the San Diego A and San Diego B datasets with more bands, the detection accuracy of hyperspectral targets achieved when there is no overlap between adjacent spectral patches is much lower than the detection accuracy when there is an overlap between adjacent spectral patches. The best detection accuracy was obtained at a length of 2 for the nonoverlapping parts between adjacent spectral patches. However, as the overlap between adjacent spectral patches gradually increases, the length of the embedded spectral sequence would also increase accordingly, leading to an increased time consumption. For the PaviaC and MUUFL Gulfport datasets with fewer bands, the detection accuracy of targets obtained with overlapping parts between adjacent spectral patches is higher than that of nonoverlapping parts between adjacent spectral patches. However, target detection accuracy is decreased when there is excessive overlap between adjacent spectral patches. The local information between adjacent spectral patches in the embedded sequence can be increased with a suitable overlap, allowing the transformer to concentrate on both the global knowledge of the spectrum and the local details in the spectrum.

2) Impact of CTFFL on Target Detection Accuracy: To investigate the effect of CTFFLs on the accuracy of HTD, we perform HTD on four HSI datasets using encoders composed of the designed CTFFL and the feedforward layer in the original transformer, respectively. All operations are point-wise in the original transformer feedforward layer, and no cross-token information can be learned [49]. The CTFFL complements the local details in the feedforward layer by adding depth-wise convolution between the two fully connected layers of the original feedforward layer. Figs. 13 and 14 represent the target detection accuracy and time consumption of HTD using feedforward and CTFFLs on four HSI datasets, respectively. It could be proved that for all HSI datasets in the experiment, the detection accuracy of HTD using the encoder composed of CTFFLs is greater than that of the encoder composed of the original feedforward layers. This demonstrates that performance improvements may result from including local detail information in the feedforward layer. However, adding depth-wise convolution between the two fully connected layers of the feedforward layer would introduce

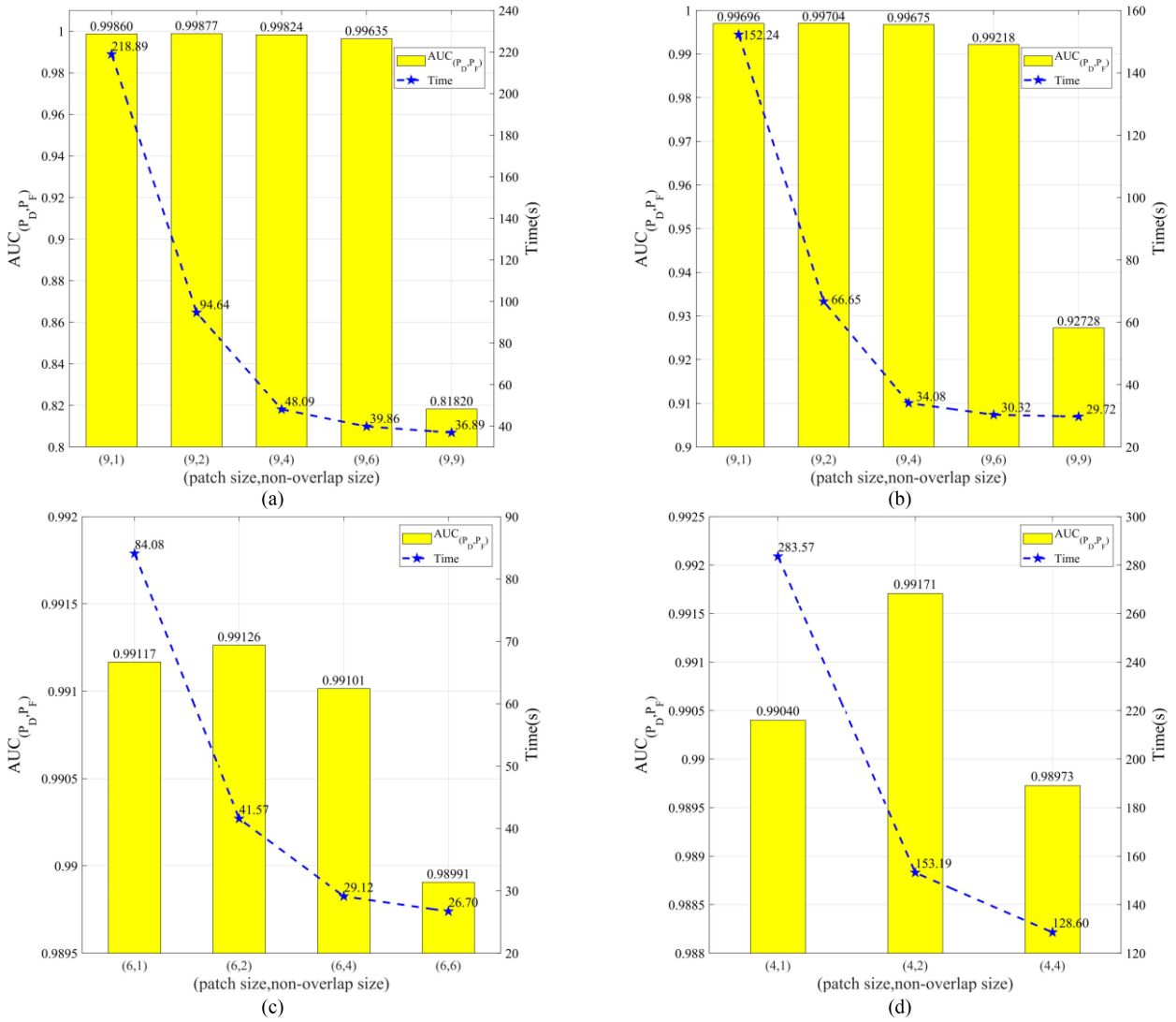

Fig. 12. Effect of different overlap sizes of adjacent spectral patches on the four HSI datasets on the accuracy of HTD. (a) San Diego A. (b) San Diego B. (c) PaviaC. (d) MUUFL Gulfport.

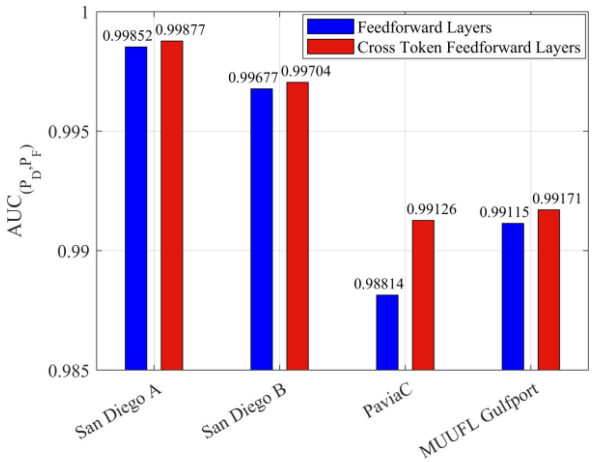

Fig. 13. Comparison of the detection performance of the CTFFL with the conventional feedforward layer for HTD on four HSI datasets.

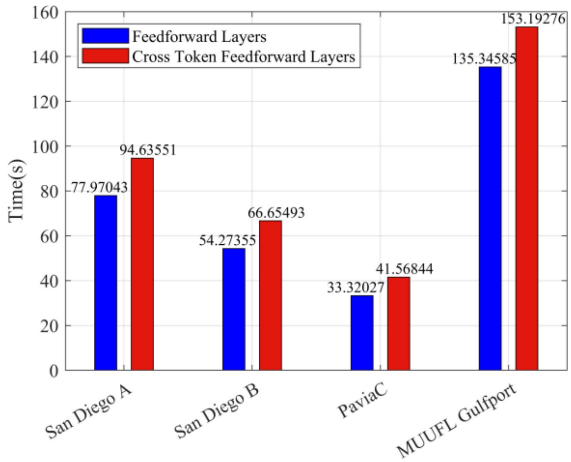

Fig. 14. Time consumption of the CTFFL versus the traditional feedforward layer on four HSI datasets.

TABLE II
TIME CONSUMPTION OF DIFFERENT METHODS FOR FOUR HYPERSPECTRAL DATASETS

Method		San Diego A	San Diego B	PaviaC	MUUFL Gulfport
CEM		0.0141	0.0121	0.0079	0.0364
OSP		0.0681	0.0485	0.0299	0.1036
CSCR		3.5560	2.2939	3.1590	751.5010
DM-BDL		4.2658	3.4805	4.1044	16.5390
BLTSC	Train	964.3975	645.8120	638.062	3274.605
	Detect	5.9217	5.8889	3.6523	3.6311
MLSN	Train	1163.754	1101.236	888.331	1099.945
	Detect	36.9408	25.8605	30.1774	163.9716
ULMMDL	Train	70.9670	51.0487	49.1577	154.2198
	Detect	0.0469	0.0416	0.2344	0.1719
MCLT	Train	94.6355	66.6549	41.5688	153.1928
	Detect	0.6437	0.4987	0.3617	0.8966

TABLE III
MODEL PARAMETERS AND COMPUTATIONS FOR FOUR DEEP LEARNING-BASED DETECTORS ON FOUR HYPERSPECTRAL DATASETS (IN MILLIONS OR THOUSANDS)

Methods	Indexes	San Diego A	San Diego B	PaviaC	MUUFL Gulfport
BLTSC	Model parameters(M)	1.57	1.57	1.50	1.47
	FLOPs (M)	17.93	17.93	6.26	4.54
MLSN	Model parameters(M)	0.05	0.05	0.05	0.05
	FLOPs (M)	9.83	9.83	6.12	3.31
ULMMDL	Model parameters(K)	7.77	7.77	4.20	2.64
	FLOPs (K)	15.53	15.53	8.39	5.28
MCLT	Model parameters(M)	0.94	0.94	0.94	0.94
	FLOPs (M)	73.73	73.73	40.09	25.69

additional learnable parameters, resulting in increased time consumption.

E. Time Consumption

Table II shows the time consumption of the seven compared methods and the proposed MCLT method together with Table III with the model complexity of deep-learning-based algorithms. Table II shows that the classical HTD methods (CEM and OSP) and representation-based HTD methods (CSCR and DM-BDL) consume much less time than the deep learning-based HTD methods. This is reasonable because deep learning-based methods require training to obtain the parameters of the network. For four deep learning-based HTD methods, the training time of the proposed method is lower than that of BLTSC and MLSN, and very close to that of ULMMDL. This is due to the transformer network used in the proposed method, which has a multiheaded self-attention mechanism that can be well parallelized on the GPU. Moreover, the training epoch of the proposed method is lower than that of the compared BLTSC and MLSN in obtaining the optimal detection results. The smaller training epoch further reduces the training time consumption of the proposed method. Once the model is well-trained, the detection efficiency depends on the detection time. The detection time of the deep learning-based detection method starts from loading the model. It ends with the detection result, as shown in Table II. The detection time of the proposed MCLT is less than the other two deep learning-based methods (BLTSC and MLSN) for the same HSI dataset. This is because the proposed method only needs to measure the similarity between the representation of the pixel spectrum to be detected and the prior target spectrum by cosine similarity at the time of detection, which can be achieved by matrix multiplication. Although the detection time consumption of MCLT is slightly more than that of ULMMDL, the detection accuracy of MCLT is higher than that of ULMMDL.

IV. CONCLUSION

In this article, a new HTD method based on unsupervised momentum contrastive learning and transformer is proposed, which can achieve excellent detection results with only one target prior spectrum. The traditional transformer has an excellent performance in focusing on spectral long-range dependencies and self-similarity, but it needs more attention to local details of the spectrum. In view of the above-mentioned problem, overlapping spectral patch embedding and CTFFLs are designed in this article to help the transformer focus on spectral local detail information. Then, a momentum encoder based on momentum update is used to extract the features of the pixel spectra for spectral discriminability learning. Finally, contrastive loss is performed for spectral discriminability learning by maximizing the similarity of positive pairs while minimizing the similarity of negative pairs. In the stage of target detection, the initial detection results are pulled up nonlinearly twice to suppress the background by using exponential-normalization, and power function-normalization operations, inspired by the function curve properties of exponential and power functions between 0 and 1. Experimental results on four real HSIs show that the proposed MCLT achieves excellent target detection and background suppression performance on HTD tasks.

REFERENCES

[1] N. M. Nasrabadi, "Hyperspectral target detection: An overview of current and future challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 34–44, Jan. 2014.

[2] C. Zhao, B. Qin, S. Feng, W. Zhu, L. Zhang, and J. Ren, "An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5546216.

[3] C. Yu et al., "Distillation-constrained prototype representation network for hyperspectral image incremental classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Jan. 2024, Art. no. 5507414.

[4] Y. Wang, H. Ma, Y. Yang, E. Zhao, M. Song, and C. Yu, "Self-supervised deep multi-level representation learning fusion-based maximum entropy subspace clustering for hyperspectral band selection," *Remote Sens.*, vol. 16, no. 2, pp. 1–18, 2024.

[5] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9646–9660, Nov. 2021.

[6] Y. Wang, X. Chen, E. Zhao, and M. Song, "Self-supervised spectral-level contrastive learning for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5510515.

[7] C. Zhao, M. Wang, S. Feng, and N. Su, "Hyperspectral target detection method based on nonlocal self-similarity and rank-1 tensor," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5500815.

[8] X. Zhao, W. Li, C. Zhao, and R. Tao, "Hyperspectral target detection based on weighted Cauchy distance graph and local adaptive collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527313.

[9] C. Zhao, W. Zhu, and S. Feng, "Superpixel guided deformable convolution network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3838–3851, May. 2022.

[10] J. Wang et al., "Hyperspectral band selection via region-aware latent features fusion based clustering," *Inf. Fusion*, vol. 79, pp. 162–173, 2022.

[11] L. Zheng, Y. Wen, W. Ren, H. Duan, J. Lin, and J. Irudayaraj, "Hyperspectral dark-field microscopy for pathogen detection based on spectral angle mapping," *Sens. Actuators B*, vol. 367, 2022, Art. no. 132042.

[12] E. Aloupogianni, M. Ishikawa, N. Kobayashi, and T. Obi, "Hyperspectral and multispectral image processing for gross-level tumor detection in skin lesions: A systematic review," *J. Biomed. Opt.*, vol. 27, no. 6, 2022, Art. no. 060901.

[13] N. M. Nasrabadi, "Regularized spectral matched filter for target recognition in hyperspectral imagery," *IEEE Signal Process. Lett.*, vol. 15, pp. 317–320, Mar. 2008.

[14] S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 1–16, Jan. 2001.

[15] J. Settle, "On constrained energy minimization and the partial unmixing of multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 3, pp. 718–721, Mar. 2002.

[16] C.-I. Chang, "Orthogonal subspace projection (OSP) revisited: A comprehensive study and analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 502–518, Mar. 2005.

[17] H. Kwon and N. M. Nasrabadi, "Kernel spectral matched filter for hyperspectral imagery," *Int. J. Comput. Vis.*, vol. 71, no. 2, pp. 127–141, 2007.

[18] H. Kwon and N. M. Nasrabadi, "Kernel adaptive subspace detector for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 2, pp. 271–275, Apr. 2006.

[19] X. Jiao and C.-I. Chang, "Kernel-based constrained energy minimization (K-CEM)," *Proc. SPIE*, vol. 6966, 2008, pp. 523–533.

[20] H. Kwon and N. M. Nasrabadi, "Kernel orthogonal subspace projection for hyperspectral signal classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 12, pp. 2952–2962, Dec. 2005.

[21] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 629–640, Jun. 2011.

[22] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, 2015.

[23] 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, Oct. 2022, Art. no. 5542814.

[24] X. Yang, B. Tu, Q. Li, J. Li, and A. Plaza, "Graph evolution-based vertex extraction for hyperspectral anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–15, Aug. 2023, doi: 10.1109/TNNLS.2023.3303273.

[25] 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.

[26] Z. Feng, J. Zhang, and J. Feng, "Spectral-spatial joint target detection of hyperspectral image based on transfer learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1770–1773.

[27] Y. Shi, J. Li, Y. Li, and Q. Du, "Sensor-independent hyperspectral target detection with semisupervised domain adaptive few-shot learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6894–6906, Aug. 2021.

[28] Y. Wang, X. Chen, F. Wang, M. Song, and C. Yu, "Meta-learning based hyperspectral target detection using Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527913.

[29] S. Mei, X. Liu, G. Zhang, and Q. Du, "Sensor-specific transfer learning for hyperspectral image processing," in *Proc. Int. Workshop Anal. Multitemporal Remote Sens. Images*, 2019, pp. 1–4.

[30] G. Zhang, S. Zhao, W. Li, Q. Du, Q. Ran, and R. Tao, "HTD-Net: A deep convolutional neural network for target detection in hyperspectral imagery," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1489.

[31] Y. Gao, Y. Feng, and X. Yu, "Hyperspectral target detection with an auxiliary generative adversarial network," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4454.

[32] D. Zhu, B. Du, and L. Zhang, "Two-stream convolutional networks for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6907–6921, Aug. 2021.

[33] W. Rao, L. Gao, Y. Qu, X. Sun, B. Zhang, and J. Chanussot, "Siamese transformer network for hyperspectral image target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5526419.

[34] Y. Shi, J. Li, Y. Zheng, B. Xi, and Y. Li, "Hyperspectral target detection with ROI feature transformation and multiscale spectral attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5071–5084, Jun. 2021.

[35] W. Xie, X. Zhang, Y. Li, K. Wang, and Q. Du, "Background learning based on target suppression constraint for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5887–5897, Sep. 2020.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[37] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 9912–9924.

[38] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[39] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21271–21284.

[40] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15745–15753.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[45] P. Wang et al., "Scaled ReLU matters for training vision transformers," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2495–2503.

[46] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[48] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[49] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10843–10852.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUS)," 2016, *arXiv:1606.08415*.

[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[53] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[54] A. Z. P. Gader, R. Close, J. Aitken, and G. Tuell, "MUUFL Gulfport hyperspectral and lidar airborne data set," Univ. of Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, 2013.
[55] X. D. A. A. Zare, "Technical report: Scene label ground truth map for MUUFL Gulfport data set," Univ. of Florida, Gainesville, FL, USA, Tech. Rep. 20170417, 2017. [Online]. Available: http://ufdc.ufl.edu/IR00009711/00001
[56] T. Cheng and B. Wang, "Decomposition model with background dictionary learning for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1872–1884, Jan. 2021.
[57] Y. Li, Y. Shi, K. Wang, B. Xi, J. Li, and P. Gamba, "Target detection with unconstrained linear mixture model and hierarchical denoising autoencoder in hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 31, pp. 1418–1432, Jan. 2022.
[58] C.-I. Chang and J. Chen, "Orthogonal subspace projection using data sphering and low-rank and sparse matrix decomposition for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8704–8722, Oct. 2021.
[59] C.-I. Chang, "An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5131–5153, Jun. 2021.
[60] L. Zhang and B. Cheng, "Fractional Fourier transform and transferred CNN based on tensor for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Apr. 2022, Art. no. 5505505.

Yulei Wang (Member, IEEE) was born in Yantai, Shandong Province, China, in 1986. She received the B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively.

She is currently an Associate Professor and Doctoral Supervisor with Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her current research interests include hyperspectral image processing and vital signs signal processing.

Dr. Wang was awarded by China Scholarship Council in 2011 as a joint Ph.D. student to study in the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County for two years. More details can be found at https://YuleiWang1.github.io/.

Xi Chen was born in Kuitun, Xinjiang Uygur Autonomous Region, China, in 2000. He received the B.E. degree in electronic information engineering from Dalian Maritime University, Dalian, China, in 2020, and the M.S. degree in information and communication engineering from Information Science and Technology College, Dalian Maritime University, Dalian, China, in 2023.

His research interests include hyperspectral target detection and deep learning.

Enyu Zhao was born in Dalian, Liaoning Province, China, in 1987. He received the Ph.D. degree in cartography and geographic information system from the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, in 2017.

From 2014 to 2016, he was a joint Ph.D. Student with Engineering Science, Computer Science and Imaging Laboratory, University of Strasbourg, Strasbourg, France. He is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. His research interests include quantitative remote sensing and hyperspectral image processing.

Chunhui Zhao received the B.S. and M.S. degrees from Harbin Engineering University, Harbin, China, in 1986 and 1989, respectively, and the Ph.D. degree from the Department of Automatic Measure and Control, Harbin Institute of Technology, Harbin, in 1998.

He was a Postdoctoral Research Fellow with the College of Underwater Acoustical Engineering, Harbin Engineering University. He is currently a Professor and a Doctoral Supervisor with the College of Information and Communication Engineering, Harbin Engineering University. His research interests include digital signal and image processing, mathematical morphology, and hyperspectral remote sensing image processing.

Dr. Zhao is a Senior Member of the Chinese Electronics Academy.

Meiping Song received the Ph.D. degree from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2006.

From 2013 to 2014, she was a Visiting Associate Research Scholar with the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore, USA. She is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. Her research interests include remote sensing and hyperspectral image processing.

Chunyan Yu received the B.S. and Ph.D. degrees in environment engineering from Dalian Maritime University, Dalian, China, in 2004 and 2012, respectively.

In 2004, she joined the College of Computer Science and Technology, Dalian Maritime University. From 2013 to 2016, she was a Postdoctoral Fellow with the Information Science and Technology College, Dalian Maritime University. From 2014 to 2015, she was a Visiting Scholar with the College of Physicians and Surgeons, Columbia University, New York City, NY, USA. She is currently an Associate Professor with the Information Science and Technology College, Dalian Maritime University. Her research interests include image segmentation, hyperspectral image classification, and pattern recognition.