remote sensing

MDPI

Article

Cross-Task Latent Feature Distillation Learning for Hyperspectral Anomaly Detection

Lei Xia [1,2] and Yinnian Liu [1,2,*]

[1] State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China; xialei22@mails.ucas.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: ynliu@mail.sitp.ac.cn

Highlights

What are the main findings?

- A cross-task latent feature distillation learning (CLFDL) framework is developed for hyperspectral anomaly detection.
- Classification knowledge, GMM-based statistical modeling, and reconstruction constraints are jointly exploited to extract valuable potential features.

What are the implications of the main findings?

- CLFDL exhibits outstanding anomaly detection capability across diverse spaceborne and airborne scenes after a one-time training.
- Joint multi-objective constraints enhance anomaly-background separability and detection accuracy in complex scenes.

Abstract

Hyperspectral anomaly detection (HAD) aims to identify Earth's surface anomalies without prior knowledge. However, most existing unsupervised HAD methods rely on training data roughly partitioned into background and anomalies, suffering from severe class imbalance and yielding feature representations heavily tied to scene-specific distributions. In contrast, supervised hyperspectral classification networks can extract well-organized spectral–spatial representations from diverse land-cover types; yet, this valuable knowledge remains underutilized in HAD due to the absence of labeled anomaly samples. To circumvent this limitation, this paper proposes a novel knowledge transfer paradigm from classification tasks to anomaly detection. The proposed framework employs a pretrained convolutional enhanced Transformer (CET) classification network as the teacher model, transferring its rich spectral–spatial context encodings via feature-level distillation. Concurrently, by synergizing latent feature alignment, Gaussian mixture model (GMM) priors, and reconstruction objectives, the student network is equipped with compact background modeling and highly discriminative representation abilities, which provide robust support for feature-space anomaly detection. Extensive experiments on spaceborne and airborne hyperspectral datasets demonstrate that our method substantially improves detection accuracy against complex terrestrial backgrounds with fast inference, while exhibiting superior cross-scene generalization.

Keywords: hyperspectral image (HSI); anomaly detection; cross-task feature learning; knowledge distillation

Academic Editor: Salah Bourennane

Received: 27 September 2026
Revised: 23 October 2026
Accepted: 26 October 2026
Published: 29 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With nanometer-scale spectral resolution across hundreds of continuous bands, HSIs support fine-grained diagnosis and identification of the physicochemical properties of land covers and are indispensable in geological exploration, environmental monitoring, and national security [1,2]. Among various hyperspectral processing tasks, HAD seeks to locate rare anomalous pixels exhibiting distinct spectral discrepancies from the dominant background without prior knowledge of target spectra or spatial distributions [3]. Nevertheless, real-world anomalies, such as artificial facilities, camouflaged targets, and minute pollutants, are characterized by low occurrence probabilities, small spatial size, and unpredictability. Therefore, the target signal is susceptible to being swamped by complex backgrounds, posing severe challenges to subtle feature extraction and robust detection [4].

Early HAD algorithms primarily leveraged statistical modeling of background features. As the canonical benchmark, the Reed-Xiaoli (RX) detector [5] calculates the Mahalanobis distance between candidate pixels and the estimated background. To accommodate nonlinearly distributed hyperspectral data, the kernel RX algorithm [6] maps pixels into an implicit high-dimensional feature space; however, this extension entails high computational complexity. Addressing the background contamination issue, weighted RX [7] assigns smaller weights to potential anomaly pixels and larger weights to background candidates, thereby achieving a more accurate estimation of the underlying background statistics. Kong et al. [8] utilized fractional Fourier entropy via the fractional Fourier transform to attenuate image noise and enlarge anomaly-background divergence. However, statistical detectors heavily rely on precise background characterization. In complex scenes with heterogeneous land-cover classes, simplistic or uniform background assumptions fail to hold. And biased estimations of background mean and covariance metrics readily lead to degraded detection performance and surging false alarms.

Representation learning provides flexible background characterization by mining intrinsic low-dimensional structures in hyperspectral data [9]. Zhao et al. [10] estimated anomalies via atom usage frequency and leveraged feature transformations with iterative reweighting to boost dictionary discriminability. Li et al. [11] introduced the collaborative representation detector, showing that l_2-norm regularization effectively models spatial correlations among background pixels and drastically reduces computation through an analytical closed-form solution. To tackle outlier pollution and insufficient spatial awareness in traditional dictionaries, Su et al. [12] incorporated principal component analysis (PCA) alongside background purification into a localized collaborative framework. Nevertheless, such local representation schemes are predominantly confined to small spatial neighborhoods and cannot capture global spatial topological contexts across large-scale scenes. Xiao et al. [13] developed a robust low-rank tensor and sparse representation that utilized tensor nuclear norms and sparsity constraints to jointly preserve global geometry and local spatial details, and further guided dictionary construction with a sparse saliency prior. Tu et al. [14] leveraged density peak clustering to drive a pure local–global collaborative dictionary-learning scheme, filtering outlier atoms through local neighborhoods while capturing globally expressive bases. Despite these representation-based methods improving decomposition capability in highly heterogeneous scenes, their underlying shallow and linear assumptions remain inadequate for characterizing complex nonlinear mixing effects. Additionally, the intensive iterative optimization required for dictionary learning drastically limits processing efficiency and deep semantic representation.

Deep learning provides powerful avenues for HAD owing to its superior nonlinear feature extraction and joint spectral–spatial representation capabilities. Existing methods commonly employ deep neural networks, such as generative adversarial networks (GANs) or autoencoders, for background reconstruction and regard pixels with large

https://doi.org/10.3390/rs18213724

reconstruction errors as anomalies [15]. For instance, Shanmugam et al. [16] proposed a dual-discriminator conditional GAN optimized via a hybrid manta-ray foraging and volcanic eruption optimization strategy, which generates realistic spatial–spectral background images to effectively decouple and identify anomalous pixels with high fidelity. Bati et al. [17] pioneered the autoencoder (AE) for HAD by encoding and decoding HSIs into sparse representations and assessing anomalies through reconstruction error. Hu et al. [18] devised a low-rank sparse AE architecture that integrates traditional dictionary representation with deep learning. This framework exploited deep AE to capture intricate spectral–spatial nuances while introducing a dictionary-learning mechanism that endows latent features with explicit physical interpretability. Despite these successes, the excessive nonlinear fitting capacity of AEs frequently leads to the undesired reconstruction of anomalies alongside the background, which severely diminishes the discriminability of reconstruction residuals and undermines anomaly separation. To mitigate this issue, several variants have imposed tailored constraints on both background and anomalies to guide HSI reconstruction. Lin et al. [19] designed a dynamic low-rank and sparse prior-constrained AE framework, where a low-rank prior refined background reconstruction while a sparse prior promoted anomaly isolation. The two AEs shared an encoder and were jointly trained in an end-to-end manner to enhance detection efficacy. Fu et al. [20] adopted a pretrained CNN denoiser as a plug-and-play prior in a low-rank representation framework and constructed a relatively pure background dictionary through clustering, thereby preventing rare pixels from being excessively suppressed by conventional low-rank denoising methods. Zhuang et al. [21] integrated explicit low-rank subspace representation, spatial self-similarity, and column-wise collaborative sparsity into a robust processing framework, enabling the model to suppress noise while explicitly preserving rare pixels and avoiding their excessive attenuation by conventional low-rank denoising methods. The incorporation of these priors effectively improves the accuracy of background modeling. However, considering the uncertainty and difficulty of obtaining prior information, some researchers have explored alternative research directions.

Taking a self-supervised masking perspective, Wang et al. [22] developed a blind-spot reconstruction network that masked target center pixels within local contexts and employed guard windows to insulate the background from anomaly interference. Huo et al. [23] used multidirectional masked convolutions to weaken the spatial correlation and reconstruction ability of anomaly pixels, while local and global dual-stream branches characterize fine-grained background features and global spectral–spatial context, respectively, thereby improving the modeling of complex backgrounds. Additionally, Gan et al. [24] used a spectral branch and a multi-scale spatial branch to extract complementary information and employed bidirectional spectral–spatial cross-attention and a multi-scale gated refinement module to strengthen cross-domain information interaction and reduce interference from complex backgrounds. Tu et al. [25] formulated a multi-scale autoencoder suppression framework that weakens anomalous responses via a self-attention module and adopted an inference-score-based dynamic masking constraint to restrict loss optimization to background pixels, thereby effectively blocking the reconstruction of anomalies.

Despite remarkable advancements achieved by deep HAD methods, the majority follow a scene-specific optimization paradigm that requires training on unlabeled pixels from the target scene before inference. Although this design adapts features to the background characteristics of the current image, it inevitably couples the learned model parameters with scene-dependent data distributions. Consequently, a model optimized for a particular hyperspectral scene tends to generalize poorly to another scene with distinct spatial–spectral attributes, which severely constrains their applicability in large-scale, continuously acquired hyperspectral remote sensing scenarios. More importantly, upon

encountering a new scene, the network generally requires additional rounds of iterative optimization before anomaly detection, resulting in substantial training overhead and prolonged detection latency.

Given that conventional unsupervised reconstruction models necessitate substantial computational resources and time for parameter re-fitting when confronted with unseen scenes, and are prone to overfitting the current scene and thereby losing generalizability, this paper presents a cross-task latent feature distillation learning framework (CLFDL) for HAD. Its student network is trained once on a Gaofen-5 (Gf-5) dataset and subsequently applied to other target test scenes with fixed parameters, thereby avoiding repeated per-scene optimization and supporting efficient cross-scene inference. Our main contributions are summarized as follows:

(1) A novel cross-task knowledge transfer paradigm is introduced for HAD. It repurposes discriminative spectral–spatial knowledge acquired from supervised HSI classification for unsupervised anomaly detection, providing a valuable blueprint for the collaborative modeling of diverse hyperspectral detection tasks.

(2) A latent representation learning strategy is developed. By jointly imposing teacher-student feature alignment, GMM-based statistical priors, and image reconstruction constraints, this strategy ensures compact background modeling and enhanced anomaly-background separability in the latent space.

(3) CLFDL supports cross-scene deployment with efficient inference. The student network is trained once on a Gf-5 scene and then applied to the remaining target test scenes with fixed parameters, without retraining for each individual scene. Comprehensive experiments on multiple spaceborne and airborne hyperspectral datasets demonstrate that this protocol avoids repeated test-scene optimization, yielding substantial accuracy gains while maintaining highly efficient inference.

2. Methodology

2.1. Overview of the Proposed Framework

Figure 1 illustrates the overall architecture of our proposed CLFDL for HAD. Concretely, we employ a convolution-enhanced Transformer (CET) [26] pretrained on the annotated Gf-5 classification dataset as the teacher model T. Its classification head is removed and its encoder remains frozen during HAD training; the teacher provides auxiliary feature constraints rather than class predictions. The student network S inherits the identical CET encoder backbone and is coupled with a lightweight convolutional decoder. For a given input HSI patch, both the teacher and student extract corresponding latent representations. Here, middle-layer relational distillation aligns normalized pixel-correlation matrices, while deep-layer cosine distillation softly aligns feature directions. Reconstruction anchors the student to the actual spectral–spatial information of the unlabeled Gf-5 training scene, and the GMM objective models its multimodal background distribution in the latent space. During inference, only the trained student model is retained for forward computation, where the ultimate anomaly score is calculated by fusing the intra-class Mahalanobis distance in the latent space with the pixel-space reconstruction residual. The following will provide a detailed introduction to these contents.

https://doi.org/10.3390/rs18213724

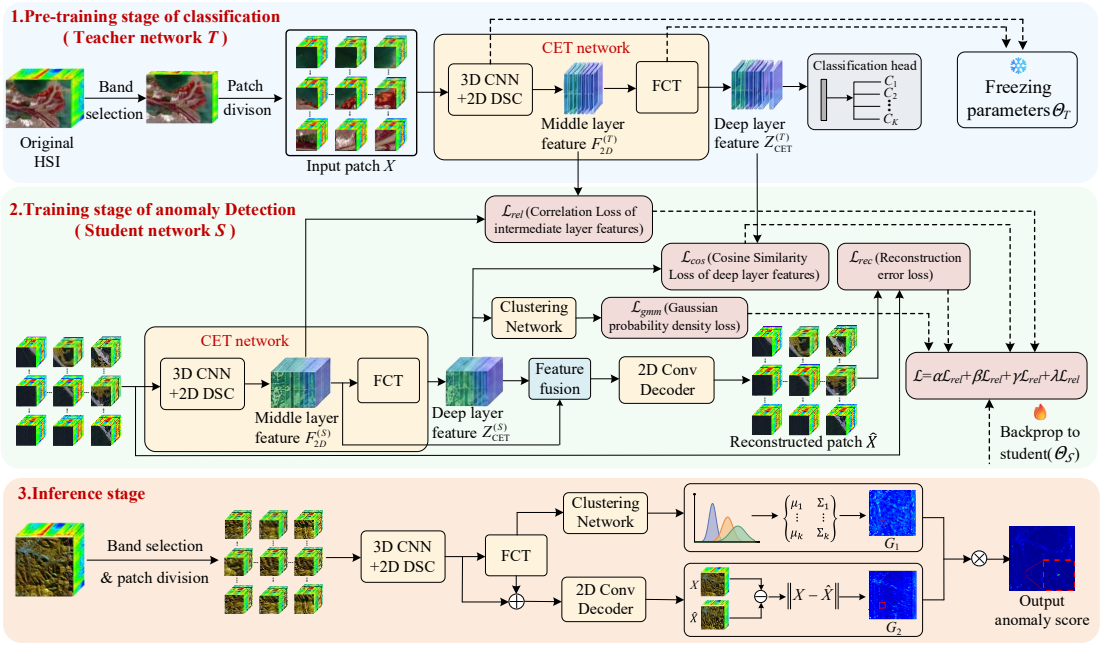


Figure 1. The implementation framework of the proposed CLFDL for HAD.

2.2. Classification-Pretrained Teacher Network

The teacher network adopts the CET network, which fully exploits the complementary strengths of CNNs and Transformers—CNNs excel at extracting local spatial information such as textures and edges, whereas Transformers are adept at capturing long-range semantic information in global spectral sequences. CET can progressively extract intermediate and deep latent features to construct a multi-scale representation hierarchy. The detailed architecture of the CET is illustrated in Figure 2, with its specific implementation detailed below.

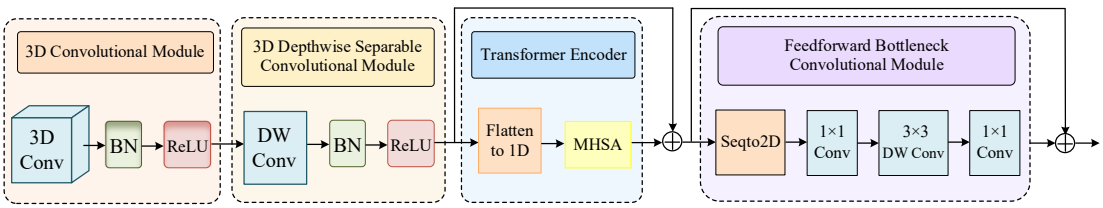


Figure 2. Schematic diagram of the CET network.

2.2.1. Local Spectral–Spatial Feature Extraction

First, a band selection (BS) algorithm is applied to compress the spectral dimension of the observed HSI $A \in \mathbb{R}^{l\times W\times b}$, yielding a dimension-reduced tensor $S \in \mathbb{R}^{l\times w\times c}$, where l and w denote the spatial height and width, respectively, b is the original number of spectral bands, and c represents the retained spectral depth. Subsequently, centered at each spatial coordinate (i, j), a local three-dimensional (3D) patch of spatial scale t × t across all c channels is cropped, resulting in a total of n volumetric patches $X \in \mathbb{R}^{t\times t\times c}$. To feed these 3D patches $\mathrm{X}_{i,j} = [\mathrm{x}_{i,j,1}, \ldots, \mathrm{x}_{i,j,c}]$ into the spectral–spatial encoding tokenizer, the input tensor is reshaped into a 5D format $E \in \mathbb{R}^{B\times 1\times t\times t\times c}$, where B designates the batch size. Thereafter, d_{out1} 3D convolutional kernels with dimensions of $k_{h0} \times k_{w0} \times k_{c0}$, named K_{3D}, are employed to jointly model the continuous spectral responses and local neighborhood geometry, producing the intermediate representation $F_{3D}^{(T)}$.

$$F_{3D}^{(T)} = \sigma\left(\mathrm{BN}(E \circledast K_{3D})\right) \in \mathbb{R}^{B\times d_{\mathrm{out1}}\times t_{h1}\times t_{w1}\times t_{c1}} \quad (1)$$

where $\circledast$ denotes the 3D convolution operator, BN represents batch normalization, $\sigma(\cdot)$ is the nonlinear activation function, and t_{h1}, t_{w1}, and t_{c1} designate the height, width, and spectral depth of the output feature map $F_{3D}^{(T)}$, respectively.

Subsequently, $F_{3D}^{(T)}$ is reshaped into a 4D tensor $F_{3D}^{\prime(T)} \in \mathbb{R}^{B\times(d_{out1}\cdot t_{c1})\times t_{h1}\times t_{w1}}$ and fed into a 2D depthwise separable convolution (DSC) module to extract higher-order deep representations $F_{2D}^{(T)}$:

$$F_{2D}^{(T)} = \sigma\left(\mathrm{BN}\left((F_{3D}^{\prime} \circledast_{dw} K_{dw}) *_{1\times 1} K_{pw}\right)\right) \in \mathbb{R}^{B\times d_{out2}\times t_{h2}\times t_{w2}} \quad (2)$$

where $\circledast_{dw}$ indicates depthwise convolution, and $K_{dw} \in \mathbb{R}^{d_{out1}\times k_{h2}\times k_{w2}}$ denotes the depthwise convolution kernel, whose channel dimension, height, and width correspond to $=\cdot c_1$, k_{h2}, and k_{w2}, respectively. $*_{1\times 1}$ represents pointwise convolution, and $K_{pw} \in \mathbb{R}^{d_{out2}\times d_{out1}\times 1\times 1}$ refers to the pointwise convolution kernel. d_{out2}, t_{h2}, and t_{w2} denote the channel number, height, and width of the resulting feature map $F_{2D}^{(T)}$.

2.2.2. Global and Local Long-to-Short Semantic Modeling

Leveraging fine-grained features extracted by CNNs can accelerate the training and convergence of the Transformer. So we flatten the feature map $F_{2D}^{(T)}$ into a sequential matrix $E_1 \in \mathbb{R}^{L\times d_{out2}}$ (where the sequence length is $L = t_{h2} \times t_{w2}$), add learnable positional embeddings $E_{pos} \in \mathbb{R}^{L\times d_{out2}}$ to form Z_0, and feed Z_0 into the multi-head self-attention (MHSA) layer:

$$Z_0 = E_1 + E_{pos} \quad (3)$$

$$Z_M = \mathrm{MSA}(\mathrm{LN}(Z_0)) + Z_0 \quad (4)$$

where $\mathrm{LN}(\cdot)$ designates layer normalization and $\mathrm{MSA}(\cdot)$ denotes the computing mechanism of MHSA [27].

Considering that conventional Transformers excel at capturing global spectral dependencies but fall short of modeling relationships among neighboring pixels, we propose a feedforward convolutional Transformer (FCT). Specifically, a lightweight bottleneck convolutional block is introduced into the multi-layer perceptron module to reinforce the localized spatial modeling capability of the feature representations.

Concretely, the output sequence of the MHSA, Z_M, is reshaped via Seqto2D into a 2D feature map Z_{2D} to incorporate spatial locality. A 1×1 convolution $W_1^{1\times1}$ is then employed to replace the first fully connected layer. To further promote information exchange among adjacent pixels, a 3×3 depthwise convolution $W_2^{3\times3}$ is introduced to capture localized contexts. Finally, another 1×1 convolution $W_3^{1\times1}$ replaces the second fully connected layer, getting the output $Z_{CET}^{(T)}$. This scheme greatly enhances local feature extraction while maintaining a low parameter footprint and computational complexity.

$$Z_{2D} = \mathrm{Seqto2D}(Z_M) \quad (5)$$

$$Z_{CET}^{(T)} = \sigma(\sigma(Z^{2D} * W_1^{1\times1}) * W_2^{3\times3}) * W_3^{1\times1} \in \mathbb{R}^{B\times d_{out3}\times t_{h3}\times t_{w3}} \quad (6)$$

where d_{out3}, t_{h3}, and t_{w3} represent the channel dimension, height, and width of the output representation $Z_{CET}^{(T)}$ produced by the FCT, respectively.

In the supervised classification task, the features extracted by the CET are fed into a linear classification head, and the parameter set θ_T of the teacher network is optimized via the standard cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{K_{cls}} y_{b,c} \log \hat{y}_{b,c} \quad (7)$$

where K_{cls} denotes the total number of ground-truth land-cover categories in the dataset, $y_{b,c} \in \{0,1\}$ is the one-hot ground-truth label of the b-th instance for class c, and $\hat{y}_{b,c}$ denotes the corresponding posterior probability estimated by the classification head.

Upon completion of training, the classification head is discarded, while the parameters θ_T of the CET encoder are preserved and frozen to serve as the teacher prior extractor for the unsupervised anomaly detection phase. By bypassing the direct transfer of explicit semantic labels and relying on latent features as the source of discriminative knowledge, this design avoids the rigid constraints imposed by the source-task semantic space on the category-unknown anomaly detection task.

Upon completion of supervised classification training, the classification head is discarded and the CET encoder parameters are frozen. The transferred information is therefore not a wetland-category decision or a class center of the trained Gf-5 dataset. It is the organization of intermediate spectral–spatial representations, including relative spectral differences, neighborhood similarity, homogeneous-region consistency, boundaries, and feature-space geometry. Earlier transfer studies show that intermediate features are generally less task-specific than output-layer decisions [28], while relational and similarity-preserving distillation can transfer sample geometry without reproducing absolute feature coordinates [29,30]. Hyperspectral graph and cross-scene studies likewise indicate that pixel relationships and spectral–spatial topology can retain useful information across scenes, although target-domain adaptation remains necessary [31–34].

2.3. Cross-Task Feature Distillation Network and Collaborative Loss Constraints

To tackle the issues of unsupervised AEs lacking class-level discriminative priors and easily learning trivial identity mappings, while mitigating the progressive attenuation of low-level geometric textures within abstract latent spaces during deep forward propagation, we design a student AE network S that integrates cross-task knowledge distillation with a cross-layer skip feature fusion mechanism. Its operational workflow is illustrated in Figure 3, with the detailed principles elaborated below.

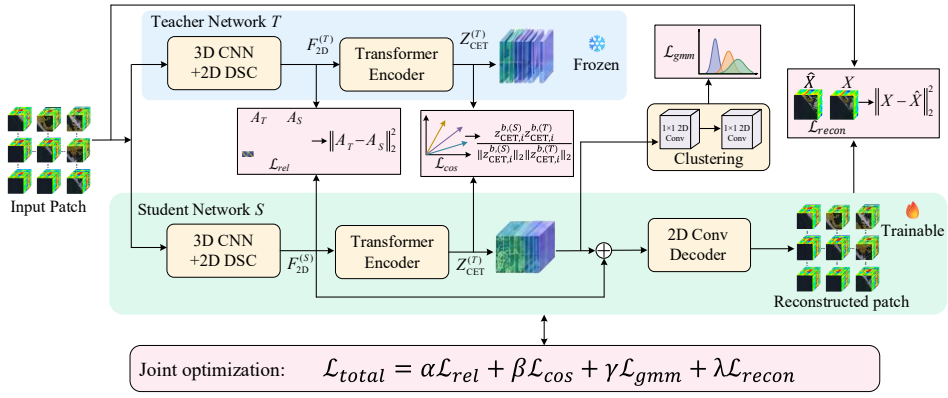


Figure 3. Training schematic of the student autoencoder network.

2.3.1. Student Autoencoder Forward Modeling and Cross-Layer Fusion

(1) Hierarchical Spectral–Spatial Feature Extraction

Local Shallow Feature Mapping: The input data is first processed by a cascaded 3D standard convolution and a 2D DSC module (as formulated in Section 2.2.1) to alleviate inter-band spectral redundancy while aggregating local spatial geometric textures, generating an intermediate localized spectral–spatial tensor $F_{2D}^{(S)}$:

$$F_{2\mathrm{D}}^{(S)} = \mathcal{H}_{\mathrm{conv}}^{(S)}\left(X;\theta_{\mathrm{conv}}^{(S)}\right) \in \mathbb{R}^{B\times d_{\mathrm{out2}}\times t_{h2}\times t_{w2}} \quad (8)$$

where $\theta_{\text{conv}}^{(S)} \subset \theta_S$ denotes the parameters of the student convolutional stream, and $\mathcal{H}_{\text{conv}}^{(S)}$ represents the combined 3D-2D convolutional functional block.

Global Deep Feature Modeling: The middle-layer representation $\mathrm{F}_{\mathrm{2D}}^{(S)}$ is subsequently serialized to feed the feedforward convolutional Transformer, which captures wide-range long-range topological dependencies across pixels over the entire receptive field, producing a highly abstract deep-layer latent feature $Z_{\mathrm{CET}}^{(S)}$:

$$Z_{\mathrm{CET}}^{(S)} = \mathcal{H}_{\mathrm{Trans}}^{(S)}\left(F_{\mathrm{2D}}^{(S)}; \theta_{\mathrm{Trans}}^{(S)}\right) \in \mathbb{R}^{B\times d_{\mathrm{out3}}\times t_{h3}\times t_{w3}} \tag{9}$$

where $\mathcal{H}_{\mathrm{Trans}}^{(S)}$ and $\theta_{\mathrm{Trans}}^{(S)} \subset \theta_S$ indicate the Transformer module and its parameters in the student network, while d_{out3} defines the deep-layer space dimension.

(2) Cross-Scale Skip Feature Concatenation

Although the deep features $\mathbf{Z}_{\mathrm{CET}}^{(S)}$ characterize abstract global manifold structures, successive convolution operations and self-attention matrix transformations tend to exert a smoothing effect on localized abrupt energy variations triggered by subtle anomalies. Conversely, the shallow features $F_{\mathrm{2D}}^{(S)}$ preserve fine-grained spatial boundaries and localized details. To this end, a cross-layer skip connection is devised to merge shallow local textures with deep global representations via channel concatenation:

$$Z_{\mathrm{fused}} = \left[F_{\mathrm{2D}}^{(S)} \,\|\, Z_{\mathrm{CET}}^{(S)}\right] \in \mathbb{R}^{B\times(d_{out2}+d_{\mathrm{out3}})\times t_{h3}\times t_{w3}} \tag{10}$$

where $[\cdot\|\cdot]$ denotes tensor concatenation along the channel dimension. This operation aggregates multi-scale spectral–spatial contexts to deliver a high-fidelity composite information stream for subsequent decoding.

(3) 2D Convolutional Reconstruction Decoder

Considering that reconstructed patches must ultimately be aggregated into a complete HSI and that transposed convolutions are prone to inducing checkerboard artifacts, we employ a two-layer 2D convolutional network to perform reverse channel expansion and spectral reconstruction:

$$F_{\mathrm{dec}} = \sigma\left(\mathrm{BN}\left(Z_{\mathrm{fused}} * K_{\mathrm{dec1}} + b_{\mathrm{dec1}}\right)\right) \in \mathbb{R}^{B\times d_{\mathrm{out2}}\times t_{h2}\times t_{w2}} \tag{11}$$

$$\hat{X} = F_{\mathrm{dec}} * K_{\mathrm{dec2}} + b_{\mathrm{dec2}} \in \mathbb{R}^{B\times c\times t\times t} \tag{12}$$

where $K_{\mathrm{dec1}} \in \mathbb{R}^{D_{\mathrm{dec}}\times(d_{\mathrm{out2}}+d_{\mathrm{out3}})\times 3\times 3}$ and $K_{\mathrm{dec2}} \in \mathbb{R}^{c\times D_{dec}\times 3\times 3}$ represent the convolutional kernel weights of the first and second decoding layers, respectively. D_{dec} denotes the output channel of the first convolutional kernel, b_{dec1} and b_{dec2} are the corresponding bias terms, and $\hat{X}$ designates the final reconstructed patch-level HSI.

2.3.2. Multi-Objective Collaborative Optimization

The joint objective assigns complementary roles to the four losses. Relational and cosine distillation provide controlled structural guidance from the frozen classification teacher. Reconstruction preserves the spectral–spatial information of the current unlabeled HAD training scene, and the GMM energy term organizes its multimodal background distribution. These objectives can compete if a single weight dominates, so the distillation constraints are deliberately soft and are balanced against target-scene reconstruction and statistical modeling.

(1) Multi-Level Feature Discriminative Knowledge Distillation Loss

Middle-layer Feature relational distillation loss ($\mathcal{L}_{\mathbf{rel}}$): To transfer local spatial topology and inter-pixel geometric correlations while circumventing the rigid overfitting caused by strict pointwise approximation, an inter-pixel correlation matrix is computed from middle-layer representations. Specifically, feature maps $F_{\mathrm{2D}}^{(T)}$ and $F_{\mathrm{2D}}^{(S)}$ are flattened along spatial dimensions into matrices $V_G, V_S \in \mathbb{R}^{B\times N_2\times d_{\mathrm{out2}}}$ $(N_2 = t_{h2} \times t_{w2})$. For each

individual sample, after normalizing the pixel feature vectors via the ℓ_2-norm, the spatial similarity correlation matrices $\mathrm{A}_T, \mathrm{A}_S \in \mathbb{R}^{N_2 \times N_2}$ for the teacher and student networks are derived as follows:

$$\mathrm{A}_T(u,v) = \frac{\mathrm{v}_{G,u}^T \mathrm{v}_{G,v}}{|\mathrm{v}_{G,u}||\mathrm{v}_{G,v}|}, \quad \mathrm{A}_S(u,v) = \frac{v_{S,u}^T v_{S,v}}{|v_{S,u}||v_{S,v}|}, \quad \forall u,v \in \{1,\dots,N_2\} \tag{13}$$

The middle-layer feature relational distillation loss is defined as the mean squared Euclidean distance between the intermediate-feature similarity correlation matrices of the teacher and student:

$$\mathcal{L}_{rel} = \frac{1}{B \cdot N_2^2} \sum_{b=1}^{B} \left\| \mathbf{A}_S^{(b)} - \mathbf{A}_T^{(b)} \right\|_2^2 \tag{14}$$

Deep Feature Cosine Similarity Distillation Loss ($\mathcal{L}_{\cos}$): In the deep semantic space, to constrain the directional alignment between the student's latent representations and the teacher's global discriminative manifold, the cosine similarity error across all pixel-wise deep feature vectors $\mathrm{z}_{\mathrm{CET},i}^{(S)}, \mathrm{z}_{\mathrm{CET},i}^{(T)} \in \mathbb{R}^{d_{\mathrm{out3}}}$ is computed as follows:

$$\mathcal{L}_{cos} = 1 - \frac{1}{B \cdot N_3} \sum_{b=1}^{B} \sum_{i=1}^{N_3} \frac{z_{\mathrm{CET},i}^{b,(S)} z_{\mathrm{CET},i}^{b,(T)}}{\|z_{\mathrm{CET},i}^{b,(S)}\|_2 \|z_{\mathrm{CET},i}^{b,(T)}\|_2} \tag{15}$$

where the architectural parameters are set as $t_{h3} = t_{h2}$ and $t_{w3} = t_{w2}$ to ensure spatial dimensional consistency between middle-layer and deep-layer features, which means $N_3 = t_{h3} \times t_{w3} = N_2$. By eliminating disturbances caused by feature magnitude variations, the cosine similarity strictly drives the student's deep features to align directionally with the categorical manifolds learned during supervised pretraining.

The rationale for using relational distillation loss for middle-layer features and cosine-similarity-based distillation loss for deep features is detailed in Appendix A.

(2) GMM Statistical Constraints in Deep Latent Spaces

To further regularize the multi-modal distribution of background representations within the deep-layer latent space, we deploy a GMM with K mixture components to model the student's deep feature representations $\mathrm{Z}_{\mathrm{CET}}^{(S)}$. A lightweight multi-layer perceptron consisting of two fully connected layers is devised to predict the posterior probability distribution $\gamma_i = [\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{iK}]^T \in \mathbb{R}^K$ of each indicating the component membership of each pixel, where $i \in \{1,2,\dots,N_3\}$ denotes the i-th pixel in the deep latent space:

$$\gamma_i = \mathrm{Softmax}\left(\mathrm{W}_{l2}\sigma\left(\mathrm{W}_{l1} Z_{\mathrm{CET}}^{(S)} + b_{l1}\right) + b_{l2}\right) \tag{16}$$

where W_{l1} and b_{l1} denote the weight matrix and bias of the first fully connected layer, respectively, while W_{l2} and b_{l2} represent those of the second fully connected layer.

Based on the current batch, the mixture probability $\hat{\phi}_k$, mean vector $\hat{\mu}_k$, and covariance matrix $\hat{\Sigma}_k$ of the k-th Gaussian component can be dynamically evaluated as [35]:

$$\hat{\phi}_k = \frac{1}{N_2} \sum_{i=1}^{N_3} \gamma_{ik} \tag{17}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N_3} \gamma_{ik} Z_{\mathrm{CET},i}^{(S)}}{\sum_{i=1}^{N_3} \gamma_{ik}} \tag{18}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{N_3} \gamma_{ik} \left(Z_{\mathrm{CET},i}^{(S)} - \hat{\mu}_k\right)\left(Z_{\mathrm{CET},i}^{(S)} - \hat{\mu}_k\right)^T}{\sum_{i=1}^{N} \gamma_{ik}} \tag{19}$$

The convergence of the multi-modal background manifold is accomplished by minimizing the sample energy via the negative log-likelihood function:

$$\mathcal{L}_{gmm} = -log \sum_{k=1}^{K} \frac{\hat{\phi}_k}{\sqrt{(2\pi)^{d_{\mathrm{out3}}} |\hat{\Sigma}_k|}} e^{\left[-\frac{1}{2}\left(Z_{\mathrm{CET},i}^{(S)} - \hat{\mu}_k\right)^T \hat{\Sigma}_k^{-1} \left(Z_{\mathrm{CET},i}^{(S)} - \hat{\mu}_k\right)\right]} \tag{20}$$

(3) Spectral–Spatial Reconstruction Loss

To establish a faithful nonlinear inverse mapping from the latent feature space back to the original observation space, the reconstruction error is employed to assess the fidelity between the reconstructed patch $\hat{X}$ and the raw input patch X:

$$\mathcal{L}_{recon} = \frac{1}{B}\sum_{b=1}^{B}\left\|X^{(b)} - \hat{X}^{(b)}\right\|_2^2 \tag{21}$$

(4) Overall Joint-Optimization Objective

The entire learnable parameter set θ_S of the student network is updated via backpropagation by jointly minimizing the overall loss function:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{rel} + \beta\mathcal{L}_{cos} + \gamma\mathcal{L}_{gmm} + \lambda\mathcal{L}_{recon} \tag{22}$$

where α, β, γ, and λ act as trade-off factors balancing the individual constraints.

2.4. Inference and Anomaly Score Generation

During the inference phase, the student network trained on the Gf-5 Huanghekou dataset remains fixed and is used for feedforward evaluation of each target test scene. No retraining or parameter reoptimization is conducted for each test scene. To fully exploit the discriminative features in the deep space domain and the reconstruction distortions in the image domain, we design a dual-domain score that couples latent intra-class Mahalanobis distances with original-space pixel reconstruction residuals.

As local homogeneous backgrounds provide stronger anomaly responsiveness than global estimators [36,37], we first execute an intra-class anomaly detection strategy in the deep latent space. The test sample is fed into the student encoder to extract the deep latent feature $\mathrm{Z}_{\mathrm{CET},m}^{(S)}$ for the m-th patch, along with the membership probability $\gamma_{i,m}$ for the i-th pixel over the GMM. The cluster label for each pixel is then assigned using the maximum a posteriori principle:

$$q_{i,m} = \arg\max_k \gamma_{i,m}^k \tag{23}$$

As a result, the patch $\mathcal{P}_m$ is adaptively partitioned into *K* statistically consistent pixel subsets. For each cluster k, the intra-group pixel set is assembled as:

$$P_m^k = \left\{Z_{\mathrm{CET},i}^{(S)} | q_{i,m} = k\right\} \tag{24}$$

Defining the pixel number as $n_m^k = |P_m^k|$ in the k-th cluster, the intra-class mean vector in the latent space is calculated as:

$$\mu_m^k = \frac{1}{n_m^k}\sum_{Z_{\mathrm{CET},i}^{(S)} \in P_m^k} Z_{\mathrm{CET},i}^{(S)} \tag{25}$$

The corresponding intra-class covariance matrix is:

$$\sum_{k,m} = \frac{1}{n_m^k - 1}\sum_{Z_{\mathrm{CET},m}^{(S)} \in P_m^k}\left(Z_{\mathrm{CET},i}^{(S)} - \mu_m^k\right)\left(Z_{\mathrm{CET},i}^{(S)} - \mu_m^k\right)^T \tag{26}$$

Consequently, for each pixel i assigned to class k within the m-th patch, its latent-space RX anomaly score is formulated as follows:

$$\delta_m^k = \left(Z_{\mathrm{CET},i}^{(S)} - \mu_m^k\right)^T \sum_{k,m}^{-1}\left(Z_{\mathrm{CET},i}^{(S)} - \mu_m^k\right) \tag{27}$$

Although the localized intra-class RX detection within the deep latent space effectively amplifies responses to subtle anomalies, its constrained receptive field may inadvertently assign high anomaly scores to land-cover boundaries and structural edges. To mitigate this phenomenon, an anomaly suppression mechanism based on reconstruction errors is further incorporated to establish complementary anomaly detection cues.

Specifically, a patch-to-image aggregation operator $\mathcal{A}(\cdot)$ maps the patch-level anomaly scores δ and reconstructed patches $\hat{X}$ back to the original spatial coordinates, forming the scene-level intra-class RX response map G_1 and the reconstructed full-scene cube S_{rec}, respectively:

$$G_1 = \mathcal{A}(\{\delta_m^k\}_{m=1}^n) \tag{28}$$

$$S_{rec} = \mathcal{A}(\{\hat{X}_m\}_{m=1}^n) \tag{29}$$

where $\mathcal{A}(\cdot)$ implicitly performs an averaging fusion operation across overlapping regions of all spatial patches.

Subsequently, the reconstruction error between the original input HSI S and the reconstructed image S_{rec} forms the second anomaly metric:

$$G_2 = \|S - S_{rec}\|_2 \tag{30}$$

To reconcile dynamic score ranges, min–max normalization is applied to both the G_1 and the G_2. A Hadamard product operation then couples discriminative cues from both domains to obtain the final 2D anomaly response map.

$$S_{final}(q) = \text{Norm}\big(G_1(q)\big) \odot \text{Norm}\big(G_2(q)\big) \tag{31}$$

where $\text{Norm}(\cdot)$ maps values to the $[0,1]$ interval, and $\odot$ is the element-wise Hadamard product.

3. Experiments and Analysis

3.1. Experimental Datasets

The hyperspectral benchmarks employed in this study primarily encompass the Gf-5, AVIRIS, and ROSIS datasets, with their corresponding ground-truth maps depicted in Figure 4. Detailed specifications are outlined as follows.

Gf-5 Dataset: Acquired by the Advanced Hyperspectral Imager (AHSI) onboard the Chinese Gf-5 satellite, the AHSI sensor spans a broad spectral range from the visible and near-infrared (VNIR: 400–1000 nm) to the short-wavelength infrared (SWIR: 1000–2500 nm) with a uniform spatial resolution of 30 m. It features spectral resolutions of 5 nm (VNIR) and 10 nm (SWIR), achieving signal-to-noise ratios (SNR) of approximately 700 and 500, respectively. For the supervised classification pretraining of the teacher network, a VNIR scene acquired over the HuangHeKou (Gf5_HHK) wetland in 2018 is adopted. The region of interest contains 462 × 617 pixels across 150 bands preserved after radiometric calibration, denoising, and atmospheric correction. Supported by concurrent field survey records and auxiliary high-resolution imagery, eight distinct land-cover categories were manually annotated to establish the ground truth for classification. Meanwhile, four SWIR scenes containing aircraft targets against diverse background landscapes were collected for the HAD evaluation. Each image measures 512 × 512 pixels, with 136 spectral bands retained following water vapor absorption band removal and denoising. Specifically, the first scene predominantly features the sea surface and adjacent coastal land, designated as Gf5_1. The second scene comprises cloud-free mountainous terrain and grassland, denoted as Gf5_2. The third scene represents a maritime background veiled by thin clouds, marked as Gf5_3. The fourth HSI mainly features a cloudy forest scene, designated as Gf5_4. Considering that Gf5_1 covers water, land, and coastal ecotones—which constitute the most quintessential background elements in remote sensing imagery—it is selected for the one-time training of the student network. After the training stage, all student-network parameters are fixed, and the same trained model is directly evaluated on the other target test scenes without per-scene retraining. Crucially, the anomalies across Gf5_1 to Gf5_4 are pixel-level minute targets accounting for a mere 0.002% of the total

pixel population, posing acute challenges to detection sensitivity and anti-interference robustness.

Gulfport Dataset: This benchmark was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over an airport area in Gulfport, Mississippi, USA. The image encompasses a spatial dimension of 100 × 100 pixels with a ground sampling distance of 3.4 m. After eliminating bands severely corrupted by water vapor absorption and low signal-to-noise ratios, 191 calibrated spectral channels are retained for empirical evaluation. The dominant background comprises airport runways, concrete pavements, and grass patches. Three aircraft parked on the tarmac serve as typical man-made anomalous targets to be detected, containing a total of 60 anomalous pixels that account for 0.60% of the entire scene [38].

Pavia Dataset: Collected by the Reflective Optics System Imaging Spectrometer (ROSIS) near a beach environment in Pavia, Italy, this dataset spans diverse land covers including water bodies, bare soil, a bridge, and multiple vehicles traveling across the bridge. The raw HSI measures 1096 × 1096 pixels with a spatial resolution of 1.3 m. In our experiments, a representative subregion of 150 × 150 pixels covering 102 valid spectral bands is cropped to identify the vehicular targets on the bridge. These vehicles are designated as anomalies, comprising 60 anomaly pixels that represent 0.30% of the cropped scene [36].



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

Figure 4. False-color ground-truth images of classification and anomaly detection datasets. (**a**) Gf5_HHK. (**b**) Gf5_1. (**c**) Gf5_2. (**d**) Gf5_3. (**e**) Gf5_4. (**f**) Gulfport. (**g**) Pavia.

3.2. Evaluation Metrics

To quantitatively and qualitatively assess the performance of the proposed CLFDL, color anomaly heatmaps, receiver operating characteristic (ROC) curves, the area under the ROC curve ($AUC_{PD\text{-}PF}$), the joint area under the curve balancing target detectability and background suppressibility ($AUC_{TD\text{-}BS}$), and box plots are employed [39,40]. Specifically, color anomaly heatmaps intuitively illustrate the spatial responsiveness of diverse detectors toward anomalies along with their capability in suppressing background clutter. The ROC curve depicts the dynamic trade-off between the probability of detection (PD) and false alarm (PF). A curve approaching the upper-left corner indicates that the detector maintains a superior anomaly hit rate under low false-alarm rates. Correspondingly, a higher integral area under the PF-PD curve, $AUC_{PD\text{-}PF}$, refers to better overall detection performance. In addition, the metric $AUC_{TD\text{-}BS}$ (as formulated in (32)) evaluates the net target detection gain under cluttered background interference by calculating the integral difference between the anomaly detectability and background false-alarm curves across variable thresholds τ. Finally, box plots characterize statistical dispersions—including extreme values, interquartile ranges, and medians—and clearly present whether a clear margin of separation is achieved between anomalous targets and background clusters.

$$AUC_{TD-BS} = AUC_{(PD,\tau)} - AUC_{(PF,\tau)} \quad (32)$$

3.3. Detection Performance for Different Methods

To conduct a comprehensive evaluation, we compare CLFDL against different types of HAD methods, including RX [5], CBAD [34], TLRSR [41], DirectNet [22], GT-HAD [42], BSRegNet [43], FERD [44], and AETNet [45]. Within traditional paradigms, RX, CBAD, and TLRSR represent classical multivariate Gaussian statistical modeling, cluster-based subspace partitioning, and low-rank tensor decomposition frameworks, respectively. Among deep learning-based approaches, DirectNet introduces a blind-spot AE architecture via local sliding-window center-pixel replacement; GT-HAD relies on local patch reconstruction paired with spatial residual aggregation; and BSRegNet adopts a full-scene-input residual regression network tailored for background manifold modeling. Additionally, FERD and AETNet serve as knowledge distillation-based anomaly detection benchmarks. Among them, RX, CBAD, DirectNet, and GT-HAD use all valid spectral bands after the standard removal of invalid bands. FERD and AETNet use the 50-band inputs specified by their original implementations. To provide comparisons under the same dimensionality-reduction setting as CLFDL, TLRSR and BSRegNet use the proposed band-selection procedure and retain 21 bands, both of which incorporate dimensionality reduction into their original formulations. Other hyperparameters follow the original papers or official implementations. This comparison avoids imposing an unvalidated common dimensionality reduction on every baseline while retaining comparisons with identical 21-band inputs. Additionally, all hyperparameters for the benchmark approaches are configured strictly in accordance with their original author-recommended settings.

For our CLFDL framework, considering that the teacher and student operate in different spectral domains (VNIR and SWIR, respectively), specific channel responses cannot be transferred directly. Nevertheless, both data sources originate from the same AHSI platform, have the same spatial resolution, and lie within the solar-reflectance imaging regime. This common imaging basis means that the two types of data still share some non-wavelength-specific information [46,47], which also provides certain favorable conditions for the migration of Gf-5 VNIR to SWIR. Additionally, during the band selection phase, the continuous spectrum of each dataset is first partitioned into seven complementary nonoverlapping subspaces. Based on the band index I_f [48], the top three critical bands maximizing the joint response of information entropy and variance within each subspace are selected, yielding a common 21-band input with high information for subsequent network processing. This operation aligns dimensionality and network structure but does not establish physical wavelength correspondence. CLFDL therefore relies on relative pixel relationships, weak directional alignment, and SWIR-driven reconstruction and GMM objectives to mitigate, rather than eliminate, the spectral-domain gap. On the other hand, to balance background homogeneity against the structural robustness of target edges, the spatial patch size for CLFDL is set to 9 × 9, aligning with the optimal neighborhood receptive field validated in [22] for hyperspectral classification. Furthermore, to reduce clustering network complexity and considering that land covers within local patches are predominantly categorized into water, land and the rest, the number of mixture clusters K is set to 3.

The anomaly detection maps generated by different methods across all datasets are illustrated in Figures 5–9. It can be observed that for the wide field-of-view and highly heterogeneous Gf-5 scenes, handcrafted shallow models, represented by RX, CBAD, and TLRSR, exhibit insufficient capability in suppressing sensor noise and complex background variations. Consequently, distinct false-alarm responses are induced in non-target regions, severely compromising the discernibility of genuine anomalies. Particularly in the Gf5_1, Gf5_3, and Gf5_4 scenes, which involve both ocean–land coexistence and cloud interference, the high-response regions induced by highly heterogeneous backgrounds further undermine the reliability of conventional detection schemes, resulting in a marked

degradation in detection accuracy. In contrast, deep neural network-driven frameworks can effectively extract latent background features through multiple iterations, thereby attenuating the background response intensity. Nonetheless, methods such as DirectNet, GT-HAD, and BSRegNet are constrained by the single-image training–single-image inference mode and exhibit a strong dependence on reconstruction error loss; consequently, their performance fluctuates noticeably across scenes with locally uneven background distributions. Benefiting from knowledge distillation, FERD, AETNet, and our CLFDL inherit robust background priors from the pretrained teacher, effectively suppressing widespread background regions while enhancing anomaly peaks across diverse datasets. However, the anomaly visualization maps produced by FERD and AETNet still inevitably exhibit some conspicuous vertical stripe artifacts, which may be attributed to the relatively large input image sizes adopted during their training, leading to insufficient modeling of local details.

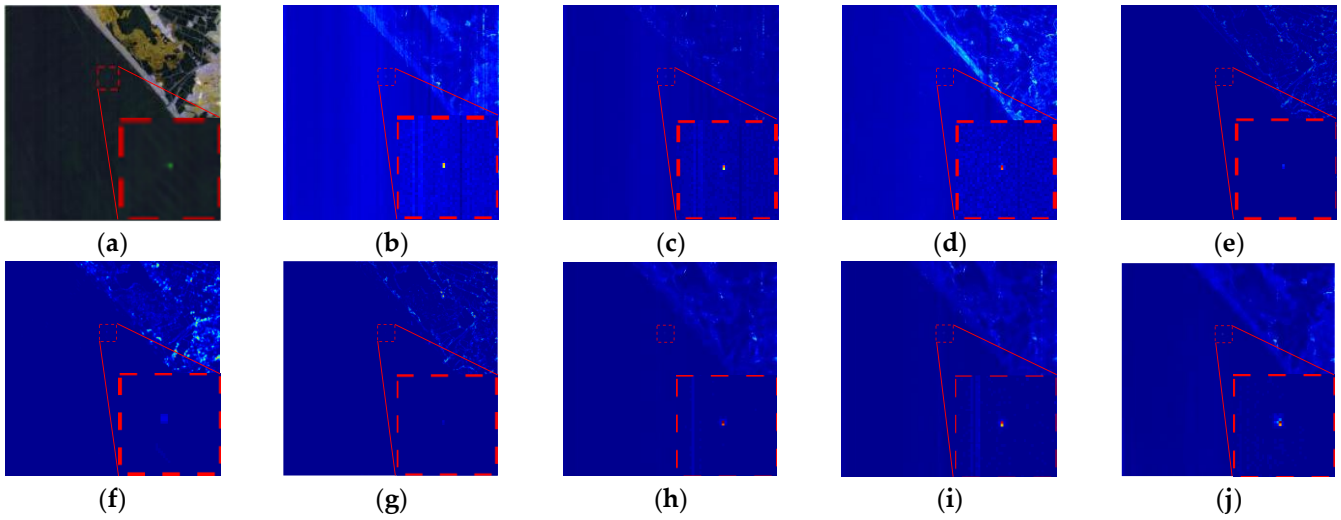


Figure 5. Anomaly heatmaps of competitive methods on the Gf5_1 dataset. (**a**) Ground truth. (**b**) RX. (**c**) CBAD. (**d**) TLRSR. (**e**) DirectNet. (**f**) GT-HAD. (**g**) BSRegNet. (**h**) FERD. (**i**) AETNet. (**j**) CLFDL.

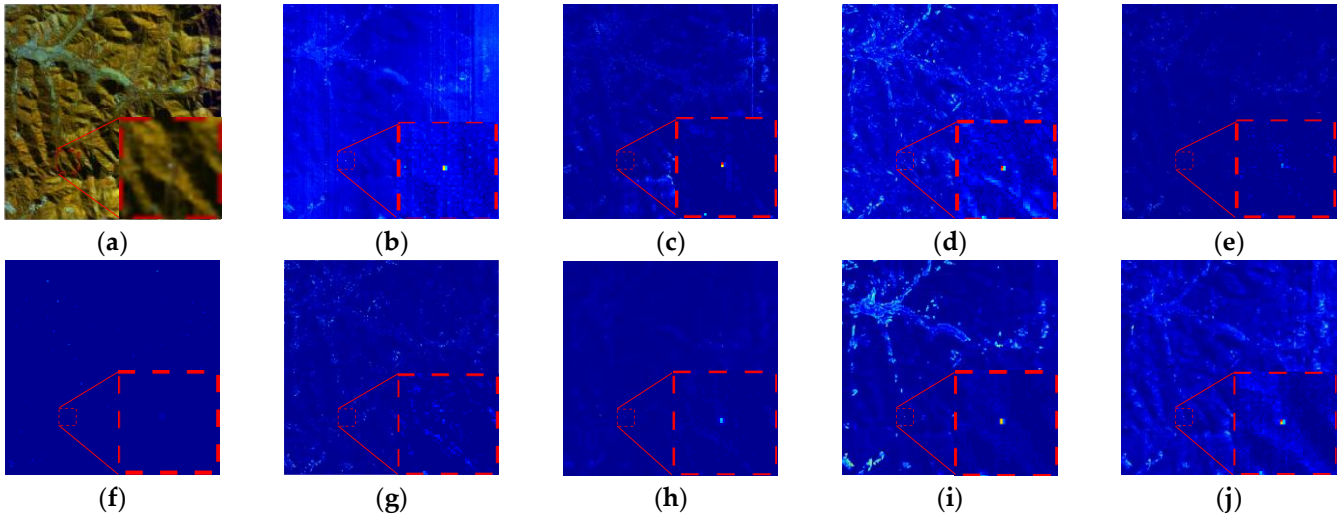


Figure 6. Anomaly heatmaps of competitive methods on the Gf5_2 dataset. (**a**) Ground truth. (**b**) RX. (**c**) CBAD. (**d**) TLRSR. (**e**) DirectNet. (**f**) GT-HAD. (**g**) BSRegNet. (**h**) FERD. (**i**) AETNet. (**j**) CLFDL.

To further assess the detection efficacy of the investigated methods, we incorporate the publicly available Gulfport and Pavia benchmark datasets, which feature a larger number of anomaly instances and higher target-to-background area ratios. On the

Gulfport dataset, although RX, CBAD, and FERD substantially attenuate background response, their detection responses toward anomalies remain rather weak, resulting in unclear anomaly spatial structures. Conversely, TLRSR, DirectNet, BSRegNet, and AETNet produce strong energy responses over anomalies, but they also induce obvious patch-like false alarms across roadways and grassland regions. Regarding the Pavia dataset, the traditional detectors, RX, CBAD, and TLRSR, produce higher false alarm rates compared to deep learning-based methods. Nonetheless, possibly owing to the relatively large target proportion, deep learning architectures are prone to mistakenly reconstruct anomalies as part of the background, leading to pronounced missed detections. In comparison, our proposed CLFDL maintains superior response stability for relatively large-scale anomalies, with anomaly regions exhibiting continuous and concentrated high responses.

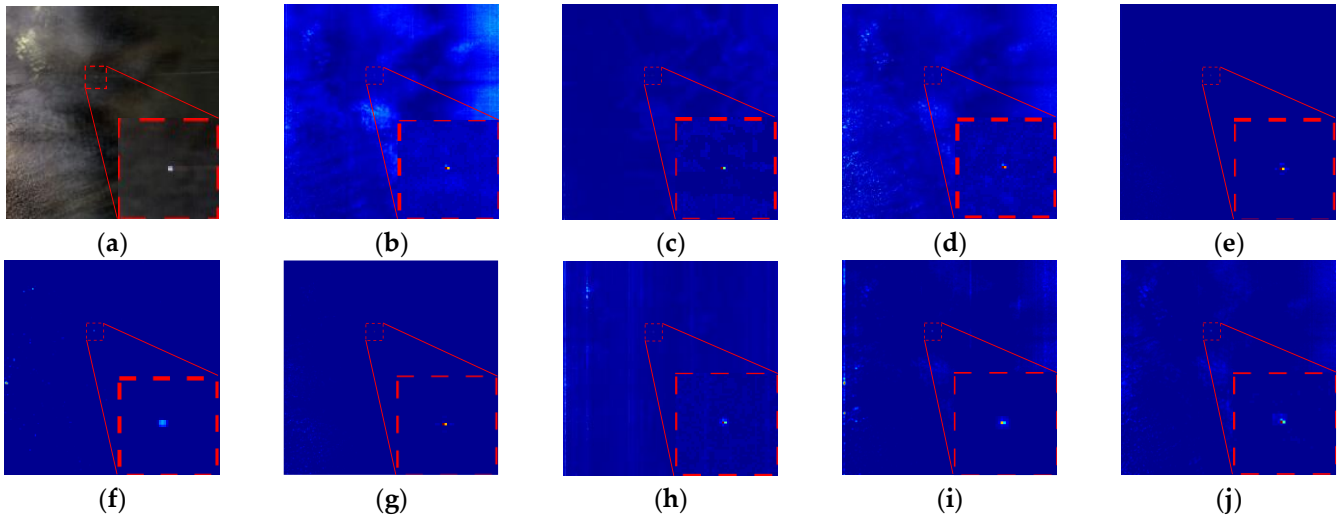


Figure 7. Anomaly heatmaps of competitive methods on the Gf5_3 dataset. (**a**) Ground truth. (**b**) RX. (**c**) CBAD. (**d**) TLRSR. (**e**) DirectNet. (**f**) GT-HAD. (**g**) BSRegNet. (**h**) FERD. (**i**) AETNet. (**j**) CLFDL.

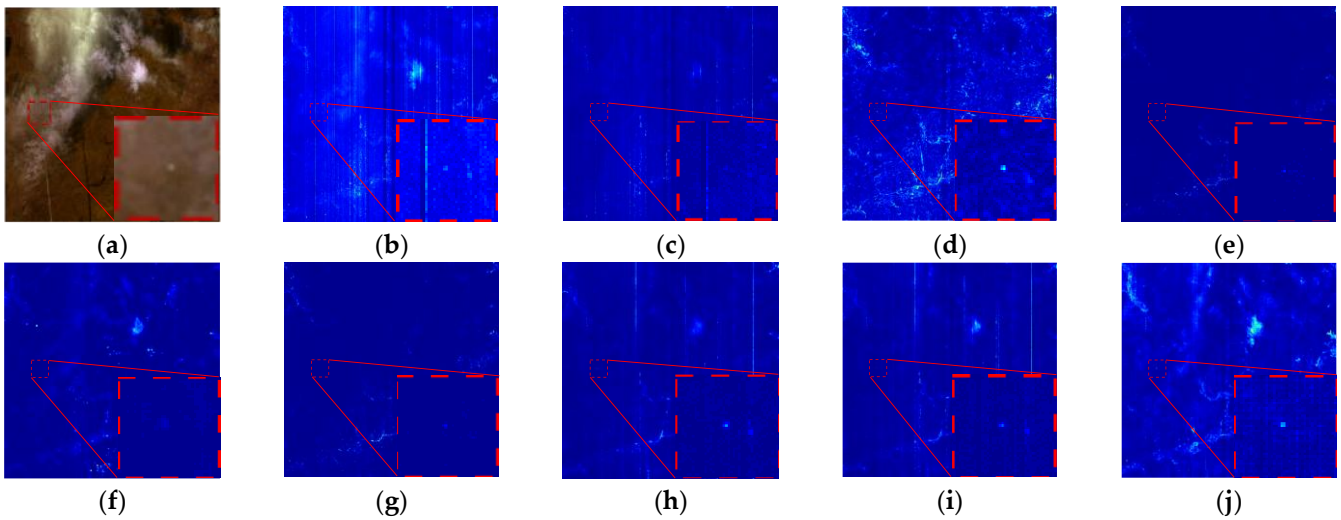
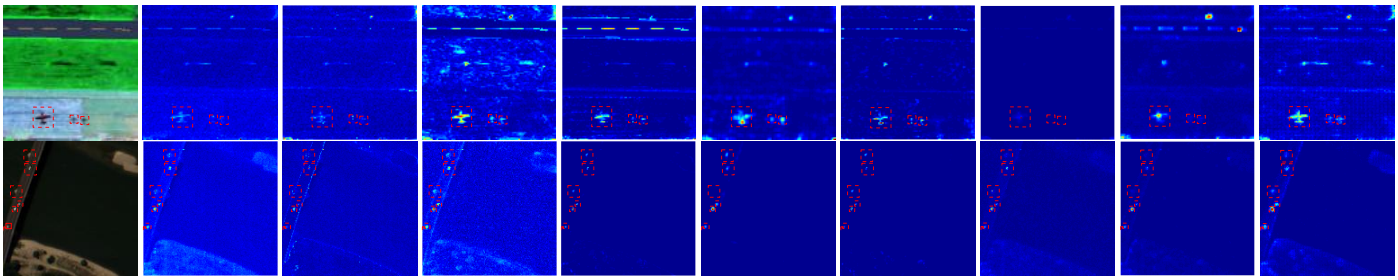


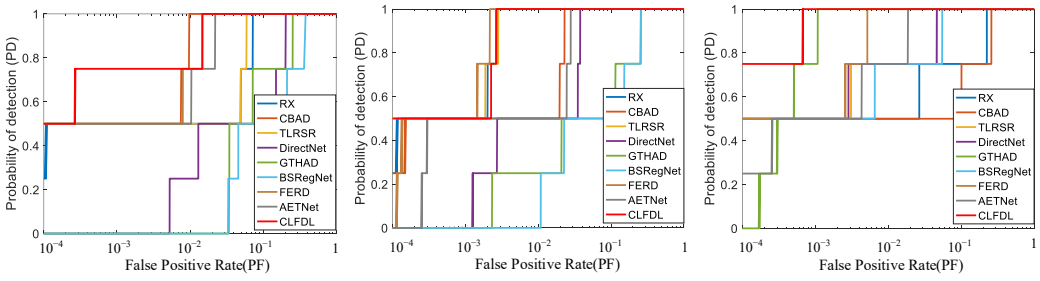
Figure 8. Anomaly heatmaps of competitive methods on the Gf5_4 dataset. (**a**) Ground truth. (**b**) RX. (**c**) CBAD. (**d**) TLRSR. (**e**) DirectNet. (**f**) GT-HAD. (**g**) BSRegNet. (**h**) FERD. (**i**) AETNet. (**j**) CLFDL.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**)

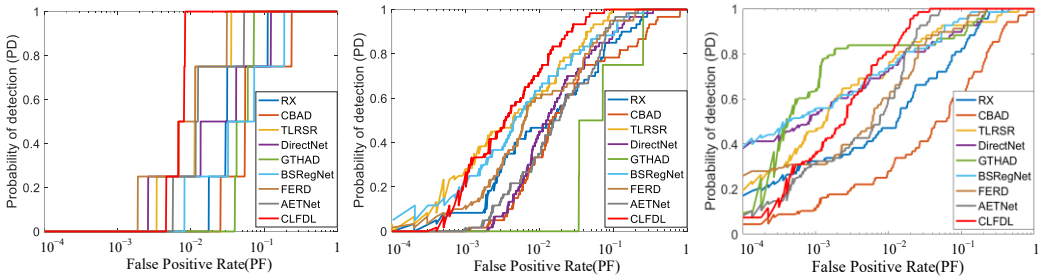
Figure 9. Anomaly heatmaps of competitive methods on the Gulfport and Pavia datasets. (**a**) Ground truth. (**b**) RX. (**c**) CBAD. (**d**) TLRSR. (**e**) DirectNet. (**f**) GT-HAD. (**g**) BSRegNet. (**h**) FERD. (**i**) AETNet. (**j**) CLFDL.

Additionally, to rigorously analyze the detection efficacy of each algorithm, we employ ROC curves to examine their dynamic discrimination trends and use the $AUC_{PD\text{-}PF}$ and $AUC_{TD\text{-}BS}$ metrics to provide fine-grained quantitative evaluation. The corresponding results are presented in Figure 10 and Table 1. Specifically, on the Gf5_1, Gf5_3, and Gf5_4 datasets, CLFDL lies closer to the top-left corner of the ROC space than most methods and achieves the peak $AUC_{PD\text{-}PF}$ score, intuitively confirming its remarkable detection performance over the comparison methods. Even on the Gf5_2 dataset, the ROC curve of CLFDL is only slightly inferior to those of FERD, RX, and TLRSR around $PF < 10^{-3}$; for $PF < 10^{-4}$, it even ranks first and attains a suboptimal $AUC_{PD\text{-}PF}$. As for the Gulfport and Pavia datasets, although conventional handcrafted models and methods such as BSRegNet lead in PD at the initial stage, the PD of CLFDL rises rapidly and surpasses that of other algorithms as PF increases, and its final $AUC_{PD\text{-}PF}$ remains second-best and best, respectively. Overall, non-distillation benchmarks exhibit significant cross-scene performance fluctuations, struggling to stably maintain an $AUC_{PD\text{-}PF}$ above 0.93. Although the distillation-based FERD and AETNet achieve improved accuracy, they still fail to reach 0.99 on all of the Gf-5, Gulfport, and Pavia datasets. In contrast, the proposed CLFDL exceeds 0.99 in $AUC_{PD\text{-}PF}$ on all evaluated datasets, demonstrating excellent discriminative accuracy and anti-interference generalization. Moreover, on $AUC_{TD\text{-}BS}$, which jointly reflects anomaly detection rate and background suppression capability, our method consistently achieves the best or second-best performance. Though TLRSR uses PCA dimensionality reduction to smooth and filter complex local backgrounds and high-frequency sensor noise, making its $AUC_{TD\text{-}BS}$ higher than that of most comparison methods, the average $AUC_{TD\text{-}BS}$ of CLFDL differs from it by only a marginal gap of about 0.013.



(**a**) (**b**) (**c**)

https://doi.org/10.3390/rs18213724



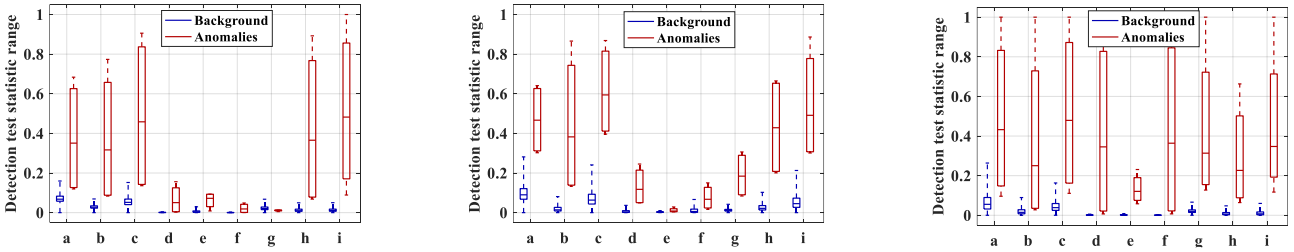
(**d**) (**e**) (**f**)

Figure 10. ROC curves of competing methods on all datasets. (**a**) Gf5_1. (**b**) Gf5_2. (**c**) Gf5_3. (**d**) Gf5_4. (**e**) Gulfport. (**f**) Pavia.

Table 1. Detection results of all methods on the test datasets (the best and second-best results per dataset are bolded and underlined, with ↑ indicating metrics for which higher values are better).

Dataset	Metrics	RX	CBAD	TLRSR	DirectNet	GT-HAD	BSRegNet	FERD	AETN	CLFDL
Gf5_1	AUC_{PD-PF} ↑	0.9694	<u>0.9956</u>	0.9727	0.9068	0.9011	0.8339	0.9943	0.9919	**0.9971**
	AUC_{TD-BS} ↑	0.3010	0.3423	0.4261	0.0583	0.0465	0.0155	<u>0.4697</u>	0.4064	**0.4994**
Gf5_2	AUC_{PD-PF} ↑	<u>0.9988</u>	0.9893	<u>0.9988</u>	0.9808	0.9020	0.8892	**0.9990**	0.9867	<u>0.9988</u>
	AUC_{TD-BS} ↑	0.3714	0.4153	**0.5366**	0.1223	0.0088	0.0589	0.1760	0.3935	<u>0.4858</u>
Gf5_3	AUC_{PD-PF} ↑	0.9370	0.9098	0.9876	0.9877	<u>0.9995</u>	0.9846	0.9981	0.9942	**0.9998**
	AUC_{TD-BS} ↑	0.4224	0.3622	**0.4706**	0.4226	0.1304	0.4319	0.4150	0.2831	<u>0.4405</u>
Gf5_4	AUC_{PD-PF} ↑	0.9522	0.9064	0.9845	0.9545	0.9462	0.9244	<u>0.9872</u>	0.9792	**0.9930**
	AUC_{TD-BS} ↑	0.0846	0.0299	**0.2056**	0.0501	0.0279	0.0373	0.1236	0.0952	<u>0.2015</u>
Gulfport	AUC_{PD-PF} ↑	0.9526	0.9141	0.9842	0.9623	**0.9932**	0.9745	0.9719	0.9654	<u>0.9907</u>
	AUC_{TD-BS} ↑	0.1004	0.0887	**0.3587**	0.2298	0.2713	0.2244	0.050	0.1705	<u>0.2998</u>
Pavia	AUC_{PD-PF} ↑	0.9538	0.8662	0.9693	0.9760	0.9779	0.9797	0.9779	<u>0.9894</u>	**0.9946**
	AUC_{TD-BS} ↑	0.1759	0.0901	**0.2482**	0.1437	0.2091	0.1225	0.1388	0.1291	<u>0.2412</u>

Figure 11 further evaluates the discriminative capability of different methods by visualizing their response distributions with box plots. It can be observed that RX, CBAD, and most competing deep learning methods have limited ability to separate background and anomalous pixels, as evidenced by the substantial overlap between their response distributions and the relatively small inter-class gaps. Although TLRSR increases the upper range of anomaly responses, it simultaneously elevates the background responses due to insufficient background constraint, thus increasing the risk of false alarms. In contrast, CLFDL consistently achieves the highest or second-highest mean anomaly response across most scenes, except for Gf5_3, while exhibiting a pronounced capability to suppress background responses. Specifically, its background responses are tightly concentrated within a low-value range, resulting in a substantially enlarged separation between the background and anomaly distributions across the benchmark datasets. These statistical characteristics provide further evidence of the superior AUC_{TD-BS} performance of CLFDL and demonstrate its ability to improve anomaly-background separability by jointly suppressing background and enhancing anomalous responses.



(**a**) (**b**) (**c**)

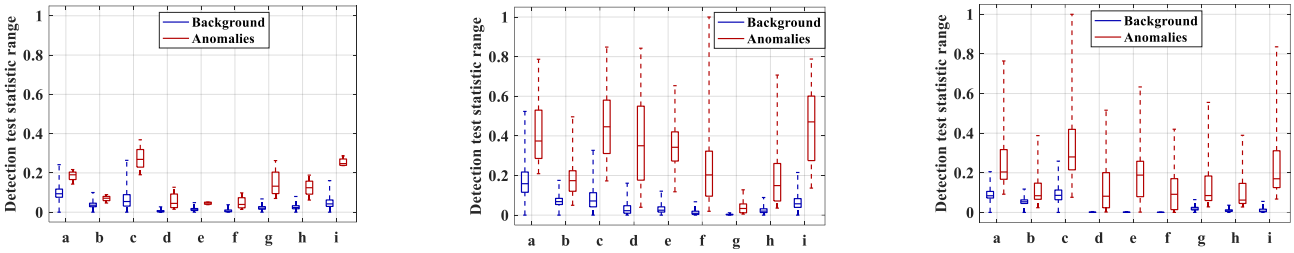


Figure 11. Separability box plots for different algorithms. (**a**) Gf5_1. (**b**) Gf5_2. (**c**) Gf5_3. (**d**) Gf5_4. (**e**) Gulfport. (**f**) Pavia. On each dataset, a: RX; b: CBAD; c: TLRSR; d: DirectNet; e: GT-HAD; f: FERD; g: SGLNet; h: AETNet; i: CLFDL.

3.4. Parameter Analysis

3.4.1. Weighting Coefficients

To investigate the sensitivity of detection performance to the weighting coefficients in the objective function, the hyperparameters α, β, γ, and λ are individually varied within $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. The corresponding $AUC_{PD\text{-}PF}$ trajectories across the four parameters are depicted in Figure 12, and the corresponding observations are discussed below.

Middle-layer features contain rich low-level details and general structural priors. For $\alpha \in [10^{-4}, 10^{-1}]$, the distillation supervision is too weak to fully model background manifolds on Gf5_1 and Gf5_4, inducing localized accuracy dips. As α increases to 1, the benefit of transferring intermediate-layer knowledge reaches a saturation point, and all datasets converge to their respective performance peaks. A further increase to $\alpha = 10$ makes the distillation constraint excessively dominant and tends to overfit the teacher's intermediate feature manifold, resulting in a slight degradation in anomaly sensitivity.

The deep latent feature distillation loss directly constrains the alignment of high-level semantic representations and is more sensitive to its weighting coefficient than the intermediate-layer distillation loss. The optimal performance is consistently obtained at $\beta = 0.01$ across all Gf5 scenes. An excessively small β provides insufficient semantic regularization, whereas $\beta \geq 0.1$ imposes an overly strict constraint on deep feature alignment. Such excessive regularization may partially suppress subtle discriminative differences between anomalous targets and background pixels, leading to performance drops of approximately 1.0% and 1.7% on Gf5_2 and Gf5_4, respectively.

The coefficient γ controls the strength of regularization imposed by the background Gaussian prior. When γ ranges from 10^{-4} to 10^{-2}, the convergence of background energy estimation is relatively slow, making it difficult for the model to establish a compact boundary for complex and mixed backgrounds. At $\gamma = 0.1$, the GMM-based energy density term provides an appropriate balance for modeling the higher-order variance of background clutter, and all datasets achieve their peak detection performance. However, an excessively large γ causes the energy penalty to dominate the backpropagation process, which may interfere with the normal reconstruction process and lead to performance degradation.

The coefficient λ controls the preservation of the spectral fidelity of the original HSI. When $\lambda \leq 0.01$, the pixel-level reconstruction constraint is relatively weak and cannot sufficiently preserve the intrinsic spectral characteristics of the input data. For example, at $\lambda = 0.001$, noticeable information distortion occurs on the low-SNR Gf5_4 scene, resulting in an $AUC_{PD\text{-}PF}$ score of only 0.96421. As λ increases to 0.1, spectral reconstruction fidelity establishes an ideal collaborative synergy with the feature distillation and GMM constraints. In contrast, when $\lambda \geq 1$, the network tends to over-reconstruct both background noise and anomalous pixels, which reduces the detectability of anomalies.

In summary, with $\alpha = 1$, $\beta = 0.01$, $\gamma = 0.1$, and $\lambda = 0.1$, CLFDL achieves high and stable anomaly scores on Gf-5 datasets. Therefore, we adopt this default hyperparameter setting consistently across all experiments in this paper.

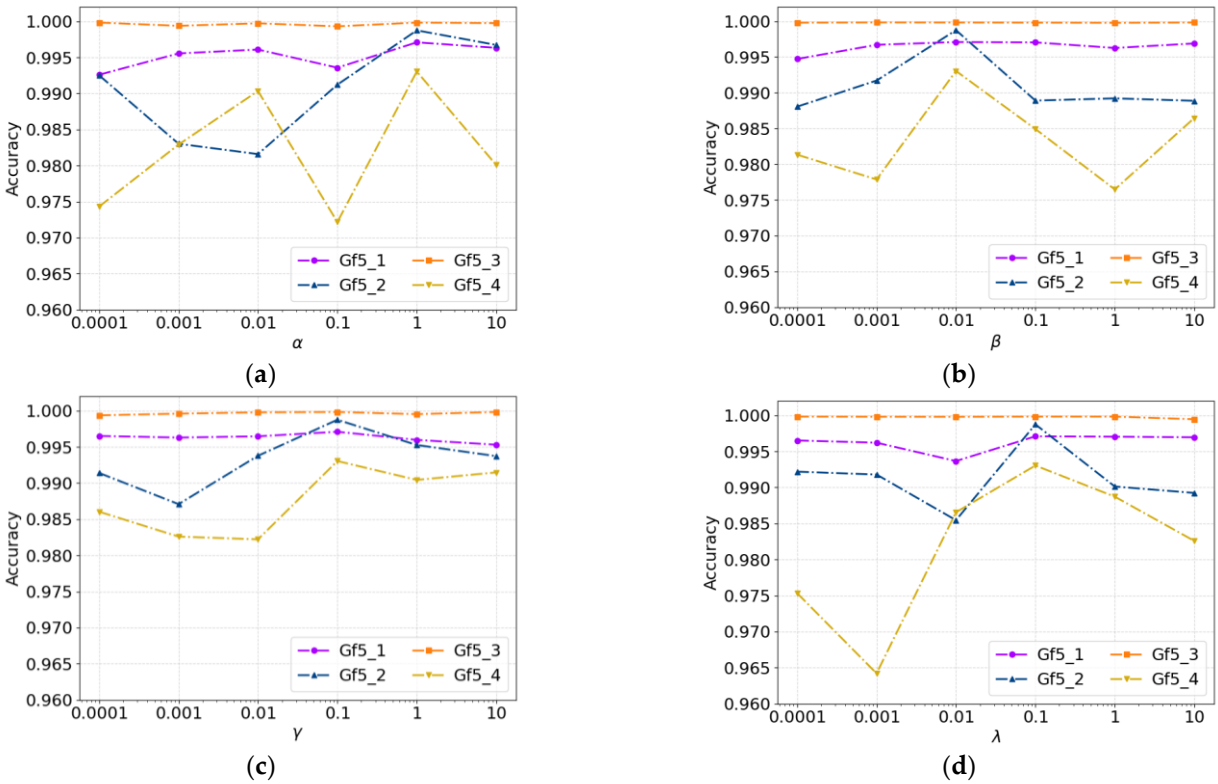


Figure 12. Parameter sensitivity analysis across different datasets. (**a**) α. (**b**) β. (**c**) γ. (**d**) λ.

3.4.2. GMM Components

To investigate the impact of GMM components K on model performance, we trained six student models on Gf5_1, each using a different value of (K). Each trained model was then directly evaluated on Gf5_2, Gf5_3, and Gf5_4 without scene-specific retraining. We used $AUC_{PD\text{-}PF}$ to assess detection performance and cross-scene generalization under different settings of (K). The results are presented in Figure 13.

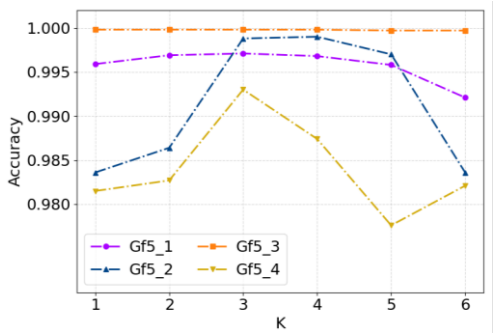


Figure 13. Effect of varying parameter K on accuracy across different test datasets.

The results show that the proposed method is generally robust to moderate variations in (K), although the optimal value varies slightly across scenes. On the source training scene, Gf5_1, the highest $AUC_{PD\text{-}PF}$ is achieved at ($K = 3$). The accuracies for $K = 2$ and $K = 4$ are 0.9969 and 0.9968, respectively, both very close to that for $K = 3$, indicating that the model exhibits good stability around the selected values.

For the cross-scene evaluations, Gf5_2 achieves its highest $AUC_{PD\text{-}PF}$ at (K = 4), although the improvement over (K = 3) is less than 0.001. A possible explanation is that its mixed mountainous and grassland background contains relatively complex background components. In contrast, Gf5_4 is more sensitive to the number of mixture components, achieving its highest $AUC_{PD\text{-}PF}$ of 0.9930 at (K = 3), with lower values under the other tested settings. Gf5_3 exhibits little sensitivity to (K), with $AUC_{PD\text{-}PF}$ remaining close to 0.9998. This behavior may be associated with its relatively simple ocean background with thin cloud cover.

These trends are consistent with the role of the GMM in our framework. When (K) is too small, particularly at (K = 1), heterogeneous background features are represented by a single, relatively coarse distribution, which may be insufficient to characterize a multimodal background comprising different land-cover types. Conversely, an excessively large (K) may partition the background within a local patch into overly fine or scene-dependent components. Such fragmentation can reduce the reliability of component-wise statistical estimates and weaken the stability of transfer to unseen scenes. We therefore use (K = 3) consistently across the datasets considered in this study, as it provides sufficient capacity to represent the dominant local background modes while avoiding unnecessary model complexity and unstable background partitioning.

3.5. Training-Scene Sensitivity

To test whether cross-scene performance depends on the student-training scene, we repeat training with each Gf-5 scene and evaluate the resulting model on all four scenes. Columns in Table 2 denote the training scene and rows denote the test scene.

The experimental results indicate that selecting Gf5_1 as the representative training scene is relatively reasonable. To evaluate the transfer performance of the different training scenes more objectively, we excluded same-scene test results and calculated the average performance of each training setting on the other three scenes. When Gf5_1, Gf5_2, Gf5_3, and Gf5_4 are used for training, the average cross-scene $AUC_{PD\text{-}PF}$ values are 0.9972, 0.9889, 0.9782, and 0.9886, respectively, and the average $AUC_{TD\text{-}BS}$ values are 0.3759, 0.3498, 0.2378, and 0.3665, respectively. Thus, training on Gf5_1 produces the highest average cross-scene performance among the four settings. Considering the results for each test scene, when Gf5_1 is used for training, $AUC_{PD\text{-}PF}$ is highest on Gf5_1, tied for the highest value on Gf5_3, and ranks second on both Gf5_2 and Gf5_4. For $AUC_{TD\text{-}BS}$, this training setting achieves the highest result on Gf5_1 and Gf5_4 and ranks second on Gf5_2. Although its result on Gf5_3 is not optimal, the land-cover composition of Gf5_3 is relatively simple, and the accuracy differences among the different settings are small. These results show that Gf5_1 does not guarantee optimal performance for every test scene and every metric, but it exhibits more stable cross-scene applicability among the four candidate scenes considered here.

Table 2. Cross-scene detection performance of CLFDLs on the Gf5 dataset under different training scenes (the best and second-best results per dataset are bolded and underlined).

Test Data \ Training Data	Metrics	Gf5_1	Gf5_2	Gf5_3	Gf5_4
Gf5_1	$AUC_{PD\text{-}PF}$	**0.9971**	0.9815	0.9885	0.9838
	$AUC_{TD\text{-}BS}$	**0.4994**	<u>0.4727</u>	0.3659	0.3975
Gf5_2	$AUC_{PD\text{-}PF}$	<u>0.9988</u>	**0.9996**	0.9890	0.9823
	$AUC_{TD\text{-}BS}$	<u>0.4858</u>	**0.4920**	0.2708	0.2506
Gf5_3	$AUC_{PD\text{-}PF}$	**0.9998**	0.9996	**0.9998**	**0.9998**
	$AUC_{TD\text{-}BS}$	0.4405	0.4344	**0.5115**	<u>0.4515</u>
Gf5_4	$AUC_{PD\text{-}PF}$	<u>0.9930</u>	0.9857	0.9462	**0.9934**
	$AUC_{TD\text{-}BS}$	**0.2015**	0.1424	0.0768	<u>0.1819</u>

One possible explanation is that Gf5_1 contains several common land-cover types, including water, land, and coastal transition zones, and therefore has a relatively diverse background composition and spatial structure. This scene composition exposes the student network to a broader range of shared spectral differences, local spatial textures, and homogeneous-region structures, which may facilitate the learning of spectral–spatial relationships that can be reused in other scenes. Thus, we consider that using Gf5_1 as training data has relative applicability in the examined scenarios. However, it cannot be assumed that the Gf5_1 dataset is a universally representative sample, which requires a larger dataset to support all unknown scenarios.

3.6. Anomaly Score Complementarity

Reconstruction-based models may also reconstruct anomalous pixels, potentially leading to missed detections. Conversely, intra-class RX anomaly detection in the latent space may be overly sensitive to local deviations, resulting in increased false alarms. This section therefore compares the detection results obtained from each method individually with those of the proposed product fusion strategy to assess the benefits of combining their complementary information.

As shown in Table 3, when only reconstruction error is used, $AUC_{PD\text{-}PF}$ on Gf5_3 is 0.9172, indicating that some anomalies in this scene may be reconstructed relatively well. On the same scene, latent-space intra-class RX achieves 0.9999, showing that the latent distribution formed through feature distillation and the GMM constraint can still identify the difference between these pixels and the corresponding background patterns. After the two scores are fused, the result on Gf5_3 reaches 0.9998. Conversely, on Gf5_2, intra-class RX alone yields 0.8164, whereas reconstruction error reaches 0.9995; the fused result remains 0.9988. These two scenes demonstrate from different perspectives that reconstruction error and latent-space intra-class RX capture different anomaly information. The lower intra-class RX score on Gf5_2 may be attributable to the complex intermixing of mountainous and grassland components in this scene, which makes the anomalies under these conditions too close to one of the Gaussian components of the surrounding background and results in insufficient intra-class Mahalanobis distances.

Table 3. Comparison of anomaly score generation strategies on the Gf-5 datasets.

Anomaly Score Generation Strategy		Gf5_1	Gf5_2	Gf5_3	Gf5_4
Reconstruction error only	$AUC_{PD\text{-}PF}$	0.9894	0.9995	0.9172	0.9762
	$AUC_{TD\text{-}BS}$	0.4890	0.6406	0.3856	0.1451
Latent intra-class RX only	$AUC_{PD\text{-}PF}$	0.9939	0.8164	0.9999	0.9922
	$AUC_{TD\text{-}BS}$	0.5515	0.2823	0.7127	0.2206
Product fusion	$AUC_{PD\text{-}PF}$	0.9971	0.9988	0.9998	0.9930
	$AUC_{TD\text{-}BS}$	0.4994	0.4858	0.4405	0.2015

Across the four Gf-5 scenes, the mean $AUC_{PD\text{-}PF}$ values obtained using reconstruction error, intra-class RX, and the combined score are 0.9707, 0.9506, and 0.9972, respectively. The minimum combined score across the four scenes is still 0.9930, whereas the minimum values for the two individual detection heads are 0.9172 and 0.8164. These results indicate that the intra-class RX supported by feature distillation and the GMM constraint is not simply appended to the reconstruction branch; rather, it provides relatively independent anomaly evidence across different scenes and thereby improves the stability of the final detection result. We also note that when the two detection heads differ substantially in their assessment of the same pixel, the lower score may weaken the final response. Thus, the current strategy can effectively reduce the risk of missed detections caused by relying

on reconstruction error alone, but it cannot guarantee that missed detections will be completely eliminated.

3.7. Ablation Study

To evaluate the contribution of each proposed module to detection performance, $AUC_{PD\text{-}PF}$ is adopted as the performance metric to quantitatively analyze the ablation effects of different modules on the Gf5 dataset. The detailed results are presented in Table 4. Specifically, Base denotes directly applying the network weights pretrained on the classification task to the encoder of the student model for anomaly detection; Base + RLoss includes the reconstruction error loss (RLoss); Base + RLoss + GLoss additionally incorporates the Gaussian probability density loss (GLoss); and Base + RLoss + GLoss + FDLoss further introduces the feature distillation loss (FDLoss) on this basis.

As shown in Table 4, the Base model achieves the lowest detection accuracy across all four scenes, indicating that although directly transferring the pretrained weights from the classification task can provide certain feature priors, the pretrained features cannot fully adapt to unsupervised anomaly detection scenarios due to the evident differences in task objectives and supervision forms between classification and anomaly detection. After adding RLoss, the accuracy on the four scenes increases by 4.98%, 2.43%, 6.86%, and 12.76%, respectively, compared with Base. This demonstrates that low-level pixel spectral reconstruction supervision injects a basic physical fidelity constraint into the AE and establishes the foundational role of spectral physical reconstruction. With the subsequent addition of GLoss, the detection scores in all scenes stabilize above 0.985. This result indicates that reconstruction error alone is insufficient to fully characterize the latent distribution of complex backgrounds; by introducing the Gaussian probability density constraint, latent features can be further modeled at the statistical distribution level, making background features more structured and compact in distribution and thereby improving the separability between anomalies and the background. Finally, after FDLoss is incorporated, the performance on the four scenes reaches 0.9971, 0.9988, 0.9998, and 0.9930, respectively, further improving over Base + RLoss + GLoss by 0.13, 1.40, 0.24, and 0.83 percentage points, respectively. Notably, although the absolute gain brought by FDLoss is not the largest relative to RLoss and GLoss, it consistently enhances detection efficacy across all four scenarios, with a pronounced improvement observed especially in the Gf5_2 scene. This indicates that feature distillation does not simply duplicate the information provided by reconstruction or probability distribution constraints, but can further preserve and transfer discriminative feature knowledge from the classification-pretrained model, enabling the student network to adapt to the HAD task while avoiding excessive deviation from the effective feature representations learned by the teacher network.

In conclusion, the proposed optimization strategy, which integrates classification-pretrained feature transfer, reconstruction constraints, probabilistic distribution regularization, and feature distillation, effectively mitigates the discrepancy between the classification and anomaly detection tasks and enhances the capability of the model to detect anomalous targets in complex hyperspectral backgrounds.

Table 4. Quantitative results of ablation experiments.

Model	Gf5_1	Gf5_2	Gf5_3	Gf5_4
base	0.9425	0.9434	0.9269	0.8658
Base + RLoss	0.9894	0.9663	0.9905	0.9763
Base + RLoss + GLoss	0.9958	0.9850	0.9974	0.9848
Base + RLoss + Gloss + FDLoss	0.9971	0.9988	0.9998	0.9930

3.8. Feature-Space Visualization

To provide a physically interpretable view of how feature distillation and Gaussian mixture model constraints improve background–anomaly separability, we present t-SNE visualizations of Gf5_1 and Gf5_3 in both the original spectral space and the deep latent feature space.

As shown in Figure 14, in the original data distribution of Gf5_1 (a), the background pixels form a broad and relatively diffuse continuous distribution, while most anomaly pixels lie inside or near the background distribution. Because the spectral variation within the background is large, the background and anomalies do not form a clear structural boundary in the original space. In the latent feature distribution (c), the background pixels are reorganized into three major clusters that are relatively compact and mutually separated. More anomaly pixels occur at the boundaries, along the extensions, or in relatively sparse regions of these clusters, increasing their distance from the high-density background centers. Gf5_4 exhibits a similar pattern. In the original data distribution (b), the background already shows a certain multimodal structure, but the different background components remain relatively diffuse; some regions are interconnected, and anomaly pixels are intermingled with the background manifold. In the latent feature distribution (d), the major background components form more compact local clusters with clearer separation between clusters. Most anomaly pixels move toward the periphery of the background clusters or toward low-density regions. Because the student network is trained only on Gf5_1, the similar change observed in Gf5_4 also indicates that the learned organization of background relationships remains effective to some extent in a scene not used for training.

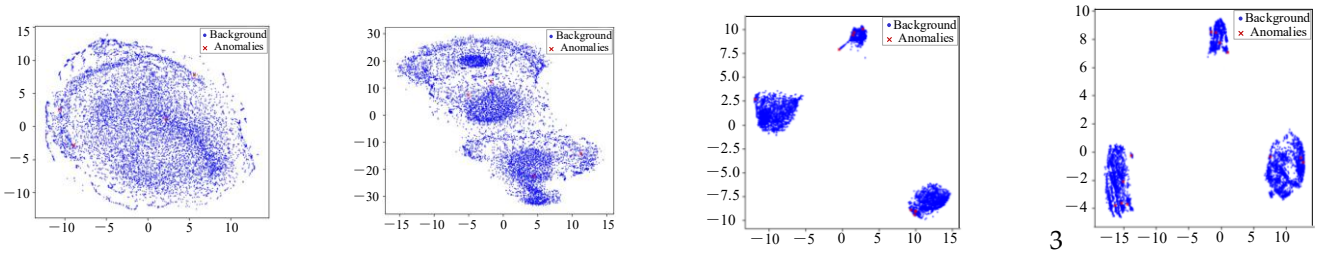


Figure 14. t-SNE distributions of the input data in the original and latent spaces: (**a**) feature distribution of Gf5_1 data in the original space; (**b**) feature distribution of Gf5_4 data in the original space; (**c**) feature distribution of Gf5_1 data in the latent space; and (**d**) feature distribution of Gf5_4 data in the latent space.

3.9. Model Efficiency and Runtime

To provide a more comprehensive assessment of model efficiency, we compare the total parameter count and FLOPs of the proposed student network with those of five deep learning baselines on the GF-5 datasets. The results are analyzed below.

As shown in Table 5, the proposed CLFDL exhibits the lowest computational complexity among all compared algorithms, requiring only 161 K parameters and 859 M FLOPs. Specifically, compared with the structurally more complex BSRegNet (670 K, 137,399 M), FERD (677 K, 115,045 M), and AETNet (890 K, 63,538 M), CLFDL achieves a substantial reduction in model parameters of 76.0–81.9% and decreases FLOPs by nearly two orders of magnitude, with an average reduction of more than 122%. Even compared with DirectNet (268 K, 70,263 M), which has a relatively small parameter count, and the next most lightweight model, GT-HAD (176 K, 910 M), our student network reduces the number of parameters by 39.9% and 8.5%, respectively, and lowers FLOPs by 98.8% and 5.6%, respectively. This substantial efficiency advantage mainly arises because the designed CET network combines the feature-learning strengths of convolutional and

Transformer models and can rapidly extract valuable deep features. In addition, the cross-task feature distillation mechanism efficiently condenses these general features into a compact student network during training. Consequently, at inference time, the model does not rely on cumbersome iterative optimization or computationally expensive feature-reconstruction modules, thereby achieving a favorable balance between detection performance and computational cost.

Table 5. Comparison of parameters and FLOPs among different deep learning methods on the Gf-5 dataset.

Methods	DirectNet	GT-HAD	BSRegNet	FERD	AETNet	CLFDL
Parameters/K	268	176	670	677	890	161
FLOPs/M	70,263	910	137,399	115,045	63,538	859

In addition, we compare the execution runtime of different comparison algorithms on the Pavia (150 × 150 pixels) and Gf-5 (512 × 512 pixels) datasets to evaluate computational overhead. All experiments were conducted using Python 3.7 on a computer equipped with a 3.0 GHz Intel Core i7 CPU, 16 GB system memory, and an NVIDIA GeForce RTX 3060 GPU with 6 GB memory. Traditional non-deep methods run on the CPU, whereas deep learning methods run on the GPU. The results are presented in Table 6. Unsupervised deep models represented by DirectNet, GT-HAD, and BSRegNet suffer from high time costs due to the scene-by-scene online training mechanism. Even on the relatively small Pavia dataset, the runtime of the above methods still exceeds 1 min; when processing large-scale Gf-5 imagery, the complete training and prediction process is further extended to approximately 30 min, limiting their rapid application in large-scale remote sensing scenarios. Although the early statistics-based RX and CBAD methods can complete prediction within 1 s on the Pavia dataset, benefiting from their direct statistical computation without iterative network training, their computational complexity associated with global matrix operations scales rapidly with expanding spatial dimensions. Consequently, their running efficiency in large-scale scenes degenerates to the same order of magnitude as distillation-based models. Among the evaluated methods, AETNet achieves the shortest runtime on the Gf-5 datasets, while our proposed CLFDL requires only approximately 10 s. It is worth noting that the runtime of CLFDL includes approximately 3 s for the band-selection-based preprocessing. Even with this preprocessing time included, CLFDL is only approximately 4.4 s slower than AETNet in terms of the overall running time. These results indicate that CLFDL substantially reduces the dependence on scene-specific retraining by reusing knowledge acquired from the classification task. This cross-task knowledge learning mechanism offers a favorable compromise between detection performance and computational cost, which provides a feasible solution for the practical deployment and scalable applications of HAD.

Table 6. Comparison of running times for all detection algorithms.

Dataset	Detectors and Running Times								
	RX	CBAD	TLRSR	DirectNet	GT-HAD	BSRegNet	FERD	AETNet	CLFDL
Pavia	0.2	0.9	5.3	280.6	76.4	255.6	13.8	3.2	3.8
Gf-5	6.4	13.0	260.5	1628.9	1810.7	590.7	18.9	6.0	10.4

4. Discussion

CLFDL uses spectral–spatial feature relationships learned through classification pretraining as an auxiliary representation prior for the anomaly detection student. The wetland classification source data, the anomaly detection training and test scenes and the AVIRIS or ROSIS test scenes differ in land-cover composition, sensor response, spatial

resolution, and spectral coverage. Consequently, transferring classification knowledge requires accounting for both task differences and domain shifts. To address these differences, CLFDL removes the teacher's classification head, provides flexible feature guidance through relational distillation and cosine similarity constraints, and adapts the student to background modeling and anomaly detection through reconstruction and GMM objectives on the HAD training scene. These mechanisms are intended to limit excessive dependence on source classification representations. Nevertheless, performance on the training scene alone does not establish cross-scene generalization, and the distillation-weight sensitivity analysis shows that overly strong teacher constraints can impair detection performance. Classification knowledge should therefore serve as a moderately weighted auxiliary prior, whose effectiveness remains dependent on task and scene differences. The proposed adaptation mechanisms mitigate the risk of negative transfer but do not guarantee its elimination.

Therefore, several limitations remain in this paper. Firstly, CLFDL is not a training-free detector. The student network still requires one-time training on Gf5_1, and its cross-scene performance depends on how well the selected training scene represents the background materials and imaging conditions encountered in unseen scenes. Secondly, although soft feature distillation, reconstruction learning, and GMM regularization mitigate cross-task and cross-domain discrepancies, they cannot completely eliminate the distribution shifts caused by different land-cover compositions, sensors, spatial resolutions, and spectral ranges. Finally, the fixed GMM component number and globally fixed loss and fusion weights may not be optimal for scenes with different levels of background complexity.

Future work will investigate multi-scene and multi-source self-supervised pre-training, physically informed spectral adaptation using wavelength and sensor-response metadata, adaptive selection of mixture components and loss weights, and broader validation across sensors, seasons, spatial resolutions, anomaly types, and onboard computing platforms.

5. Conclusions

To address the limitations of existing unsupervised HAD methods, which rely heavily on scene-specific data distributions, require repeated model training for different scenes, and often exhibit limited cross-scene generalization, this paper proposes a CLFDL framework that facilitates knowledge transfer across different hyperspectral detection tasks. CLFDL first exploits the rich discriminative spectral–spatial knowledge acquired from supervised classification to construct a teacher model, and then transfers deep representations to the anomaly detection task through feature distillation. Meanwhile, a reconstruction constraint is introduced to preserve the spectral–spatial structural information of the input data, while a GMM-based statistical prior further regularizes the background distribution in the deep latent feature space. These complementary components are jointly integrated into a unified detection framework, which simultaneously leverages pretrained knowledge, background modeling, and feature representation. Comprehensive experiments on multiple spaceborne and airborne hyperspectral datasets demonstrate the effectiveness of CLFDL in terms of detection accuracy and cross-scene adaptability. In addition, runtime evaluations show that CLFDL can perform anomaly detection on new scenes with low computational overhead. Overall, this work provides a new paradigm for feature learning in HAD and offers a potential avenue for knowledge reuse and collaborative learning across different hyperspectral remote sensing tasks.

Although CLFDL demonstrates effective cross-scene detection under the evaluated conditions, its generalization may still be influenced by the representativeness of the training scene of the student, particularly under substantial cross-sensor or spectral

distribution shifts. Future work will investigate multi-scene and multi-source pretraining, together with sensor-aware spectral adaptation and adaptive background modeling, to further improve model robustness.

Author Contributions: Conceptualization, L.X. and Y.L.; methodology, L.X.; investigation, L.X. and Y.L.; writing—original draft preparation, L.X.; writing—review and editing, Y.L.; funding acquisition, Y.L.; supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Program of the National Natural Science Foundation of China under Grant NO. 42192582, the National Key Research and Development Program of China under Grant No. 2022YFB3902000, and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDB0580000.

Data Availability Statement: The data will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Two-Level Distillation Design Analysis

The central motivation for the two-level distillation strategy is that features at different network depths have different levels of abstraction and therefore should not be constrained in exactly the same way. Middle-layer features retain relatively fine-grained spectral–spatial structures and are sensitive to local variations and task-related activation patterns. Because the teacher and student networks are designed for different learning tasks, directly requiring their feature vectors at this layer to align one by one in direction may impose an excessively strong constraint on the student network. It may also transfer activation details that are closely related to the teacher task but are not necessarily suitable for hyperspectral anomaly detection. In contrast, relational distillation constrains the relative organization or pairwise geometry among features rather than requiring the student network to reproduce the teacher representation coordinate by coordinate. It can therefore retain structural knowledge from the teacher while preserving the representational freedom needed for the student to adapt to the target task. This design motivation is consistent with the main conclusions of previous studies on relational knowledge distillation and similarity-preserving distillation. Transferring relational information avoids requiring the student network to imitate the teacher representation space point by point [29,30].

By contrast, deep features become more compact after multiple transformations and aggregate higher-level discriminative information from the teacher network. At this level, cosine similarity distillation can provide a direct directional constraint between corresponding teacher and student features. Because cosine similarity is unaffected by feature magnitude, this constraint can transfer the discriminative directions in the teacher's deep representation without requiring the student network to reproduce the absolute response magnitude of the teacher features. This property is particularly important for cross-task distillation because the student network must retain sufficient freedom to satisfy its own anomaly detection objective. Previous studies have also shown that cosine-similarity-based metrics are suitable for comparing high-dimensional feature representations [49]. Recent research on feature distillation has further indicated that features at different network depths may require different distillation strategies [50].

After the two losses are exchanged, their constraint mechanisms do not fully match the representation characteristics of the corresponding layers. Applying cosine similarity distillation to the middle layer imposes a direct directional-consistency constraint on features that still contain substantial local variation and task-related information, which may weaken the student's ability to adapt to the anomaly detection task. Conversely, applying

only relational distillation to the deep layer can preserve the relative geometry among deep samples, but it cannot directly anchor each student deep feature to the corresponding teacher feature. We therefore do not argue that relational distillation is inherently unsuitable for deep features. Rather, in the proposed cross-task architecture, it provides a less direct constraint on high-level discriminative information than sample-wise cosine directional alignment.

To verify this analysis, we compared the configuration used in this study, which combines middle-feature relational distillation and deep-feature cosine similarity distillation (denoted MFRD-DFCD), with the exchanged configuration, which combines middle-feature cosine similarity distillation and deep-feature relational distillation (denoted MFCD-DFRD). As shown in Table A1, MFRD-DFCD achieves better results on both evaluation metrics across all four Gf-5 scenes. From Gf5_1 to Gf5_4, its $AUC_{PD\text{-}PF}$ increases by 0.0016, 0.0092, 0.0002, and 0.0089, respectively, and the mean increases from 0.9922 to 0.9972. For $AUC_{TD\text{-}BS}$, which shows a more pronounced difference, the corresponding absolute gains are 0.0305, 0.1140, 0.1181, and 0.0454, respectively. The mean increases from 0.3298 to 0.4068, representing a relative improvement of approximately 23.3%.

Table A1. Accuracy results under different feature distillation configurations.

Method \ Dataset	Metrics	Gf5_1	Gf5_2	Gf5_3	Gf5_4
MFRD–DFCD	$AUC_{PD\text{-}PF}$	0.9971	0.9988	0.9998	0.9930
	$AUC_{TD\text{-}BS}$	0.4994	0.4858	0.4405	0.2015
MFCD–DFRD	$AUC_{PD\text{-}PF}$	0.9955	0.9896	0.9996	0.9841
	$AUC_{TD\text{-}BS}$	0.4689	0.3718	0.3224	0.1561

Compared with the nearly saturated $AUC_{PD\text{-}PF}$, the two configurations show a more pronounced difference in $AUC_{TD\text{-}BS}$. This result indicates that the current loss allocation mainly improves the overall balance between target detectability and background suppression rather than producing only a small effect on target-ranking performance. The improvements on Gf5_2 and Gf5_3 are especially clear, suggesting that, in scenes with more complex background variation, preserving middle-layer structural information through relational distillation while directly aligning discriminative directions at the deep layer is the more appropriate combination. These results provide direct experimental evidence for the design in which relational distillation is applied to the middle layer and cosine similarity distillation is applied to the deep layer.

References

1. Garske, S.; Evans, B.; Artlett, C.; Wong, K.C. ERX: A fast real-time anomaly detection algorithm for hyperspectral line scanning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5503617.
2. Nie, X.; Xue, Z.; Li, J. Model knowledge-prior embedded subspace learning network for hyperspectral image classification. *Pattern Recognit.* **2026**, *179*, 113587.
3. Qu, B.; Zheng, X.; Qian, X.; Lu, X. Research progress on hyperspectral anomaly detection. *Natl. Remote Sens. Bull.* **2024**, *28*, 42-54.
4. Xing, Y.; Hang, L.; Chen, Z.; Peng, R.; Chen, J.; Zou, J.; Dong, W.; Wang, Q. Real-time remote sensing for sudden surface anomalies: A review of principles and challenges. *Space Sci. Technol.* **2025**, *5*, 0362.
5. Reed, I.S.; Yu, X. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1760–1770.
6. Wu, Z.; Wang, B. Kernel-based decomposition model with total variation and sparsity regularizations via union dictionary for nonlinear hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16.
7. Guo, Q.; Zhang, B.; Ran, Q.; Gao, L.; Li, J.; Plaza, A. 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.* **2014**, *7*, 2351–2366.

8. Kong, W.; Shen, D.; Li, M.; Ma, X.; Wang, H. Hyperspectral anomaly detection based on tensor approximation with tensor double nuclear norm. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5507414.
9. Wang, X.; Wang, Y.; Mu, Z.; Zhang, Y. UFBSM: Unmixing fusion and background sparse dictionary model for hyperspectral anomaly detection. *Int. J. Remote Sens.* **2024**, *45*, 3541–3559.
10. Zhao, R.; Du, B.; Zhang, L. Hyperspectral Anomaly Detection via a Sparsity Score Estimation Framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3208–3222.
11. Li, W.; Du, Q. Collaborative representation for hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1463–1474.
12. Su, H.; Wu, Z.; Du, Q.; Du, P. Hyperspectral anomaly detection using collaborative representation with outlier removal. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2018**, *11*, 5029–5038.
13. Xiao, Q.; Zhao, L.; Chen, S.; Li, X. Robust tensor low-rank sparse representation with saliency prior for hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20.
14. Tu, B.; Yang, X.; Li, N.; Zhou, C.; He, D. Hyperspectral anomaly detection via density peak clustering. *Pattern Recognit. Lett.* **2020**, *129*, 144–149.
15. Sun, X.; Zhang, Y.; Dong, Y.; Du, B. Contrastive Self-Supervised Learning-Based Background Reconstruction for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5504312.
16. Shanmugam, P.; Amali, S.A.M.J. Dual-discriminator conditional generative adversarial network optimized with hybrid manta ray foraging optimization and volcano eruption algorithm for hyperspectral anomaly detection. *Expert. Syst. Appl.* **2024**, *238*, 122058.
17. Bati, E.; Çalışkan, A.; Koz, A.; Alatan, A.A. Hyperspectral anomaly detection method based on auto-encoder. *Proc. Image Signal Process. Remote. Sens. XXI* **2015**, *9643*, 96430N.
18. Hu, X.; Li, Z.; Luo, L.; Karimi, H.R.; Zhang, D. Dictionary trained attention constrained low rank and sparse autoencoder for hyperspectral anomaly detection. *Neural Netw.* **2025**, *181*, 106797.
19. Lin, S.; Zhang, M.; Cheng, X.; Shi, L.; Gamba, P.; Wang, H. Dynamic Low-Rank and Sparse Priors Constrained Deep Autoencoders for Hyperspectral Anomaly Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–18.
20. Fu, X.; Jia, S.; Zhuang, L.; Xu, M.; Zhou, J.; Li, Q. Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9553–9568.
21. Zhuang, L.; Gao, L.; Zhang, B.; Fu, X.; Bioucas-Dias, J.M. Hyperspectral image denoising and anomaly detection based on low-rank and sparse representations. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–17.
22. Wang, D.; Zhuang, L.; Gao, L.; Sun, X.; Zhao, X.; Plaza, A. Sliding Dual-Window-Inspired Reconstruction Network for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15.
23. Huo, Y.; Wang, S.; Wang, C.; Zhang, M.; Wang, H. Dual-stream background modeling network with anomaly suppression for hyperspectral anomaly detection. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *148*, 105233.
24. Gan, Y.; Wang, M.; Zhang, L.; Zhang, W.; Yu, T.; Wang, H. A balanced spectral–spatial cross-fusion network for hyperspectral anomaly detection. *Remote Sens.* **2026**, *18*, 2820.
25. Tu, B.; Zhou, T.; Liu, B.; He, Y.; Li, J.; Plaza, A. Multi-Scale Autoencoder Suppression Strategy for Hyperspectral Image Anomaly Detection. *IEEE Trans. Image Process.* **2025**, *34*, 5115–5130.
26. Xia, L.; Liu, Y.; Zhao, S.; Li, P.; Wang, Q.; Zhao, Y.; Chai, M. A Lightweight Spectral-Spatial Unified Feedforward Convolutional Transformer Framework for Spaceborne Hyperspectral Image Classification. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2025**, *19*, 3123–3142.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
28. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 3320–3328.
29. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3962–3971.
30. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1365–1374.
31. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978.

32. Wang, D.; Du, B.; Zhang, L. Spectral-spatial global graph reasoning for hyperspectral image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 12924–12937.
33. Zhong, C.; Zhang, J.; Wu, S.; Zhang, Y. Cross-scene deep transfer learning with spectral feature adaptation for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2861–2873.
34. Rao, W.; Qu, Y.; Gao, L.; Sun, X.; Wu, Y.; Zhang, B. Transferable network with Siamese architecture for anomaly detection in hyperspectral images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102669.
35. Xing, H.; Wu, H.; Zhang, P. Deep multi-sphere support vector data description based on disentangled representation learning. *Pattern Recognit.* **2024**, *156*, 110842.
36. Carlotto, M. A cluster-based approach for detecting man-made objects and changes in imagery. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 374–387.
37. Zhao, C.; Wang, Y.; Qi, B.; Wang, J. Global and local real-time anomaly detectors for hyperspectral remote sensing imagery. *Remote Sens.* **2015**, *7*, 3966–3985.
38. Kang, X.; Zhang, X.; Li, S.; Li, K.; Li, J.; Benediktsson, J.A. Hyperspectral anomaly detection with attribute and edge-preserving filter. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5600–5611.
39. Chang, C.I. An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5131–5153.
40. Fu, X.; Zhang, T.; Cheng, J.; Jia, S. MMR-HAD: Multiscale Mamba reconstruction network. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5516914.
41. Wang, M.; Wang, Q.; Hong, D.; Roy, S.K.; Chanussot, J. Learning tensor low-rank representation for hyperspectral anomaly detection. *IEEE Trans. Cybern.* **2023**, *53*, 679–691.
42. Lian, J.; Wang, L.; Sun, H.; Huang, H. GT-HAD: Gated Transformer for hyperspectral anomaly detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 3631–3645.
43. Hu, J.; Zheng, W.; Wang, R.; Zhao, M. A band-selected and regularized network for hyperspectral anomaly detection. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–14.
44. Jin, W.; Dang, F.; Zhu, L. Feature enhancement with reverse distillation for hyperspectral anomaly detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
45. Li, Z.; Wang, Y.; Xiao, C.; Ling, Q.; Lin, Z.; An, W. You Only Train Once: Learning a General Anomaly Enhancement Network with Random Masks for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5506718.
46. Liu, Y.N.; Sun, D.-X.; Hu, X.-N.; Ye, X.; Li, Y.-D.; Liu, S.-F.; Cao, K.-Q.; Chai, M.-Y.; Zhou, W.-Y.-N.; Zhang, J.; et al. The Advanced Hyperspectral Imager: Aboard China's GaoFen-5 satellite. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 23–32.
47. Green, R.O.; Eastwood, M.L.; Sarture, C.M.; Chrien, T.G.; Aronsson, M.; Chippendale, B.J.; Faust, J.A.; Pavri, B.E.; Chovit, C.J.; Solis, M.; et al. Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sens. Environ.* **1998**, *65*, 227–248.
48. Du, Z.; Jeong, M.K.; Kong, S.G. Band selection of hyperspectral images for automatic detection of poultry skin tumors. *IEEE Trans. Autom. Sci. Eng.* **2007**, *4*, 332–339.
49. Passalis, N.; Tefas, A. Learning deep representations with probabilistic knowledge transfer. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 283–299.
50. Yang, Z.; Li, Z.; Zeng, A.; Li, Z.; Yuan, C.; Li, Y. ViTKD: Feature-based knowledge distillation for vision transformers. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 1379–1388.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.