Hyperspectral Target Detection Based on a Background-Aware Sparse Transformer Network

Zhiwei Wang, Kun Tan, *Senior Member, IEEE*, Xue Wang, and Xiaojun Xu

***Abstract*—Hyperspectral target detection (HTD) relies on prior target spectra to locate the targets of interest within hyperspectral images (HSIs). Recently, deep learning methods have shown their potential in hyperspectral feature extraction and multiscale feature fusion. In this article, we propose a background-aware sparse transformer network (BASTNet) for HTD, to solve the problems of target sample imbalance and underutilization of global information. First, the proposed method utilizes random masking and target spectra generation strategies to establish an image-level training paradigm, constructing sufficient and balanced training samples to prompt the network to learn spatial-contextual features between the target and the background. We then introduce a Siamese sparse transformer network (S^{2}TNet) with an encoder–decoder structure to achieve fast inference for large-scene hyperspectral imagery. Specifically, S^{2}TNet consists of pyramid feature extraction, multiscale feature fusion, and a target detector, with a sparse self-attention mechanism enhancing the focus on target regions and improving the separability between target and background. Furthermore, a background-aware learning mechanism is introduced that uses a foreground and background guidance loss that attenuates the interference of background noise on the target detection. Experiments on five benchmark datasets demonstrate the superiority and applicability of the proposed BASTNet method, showing that it outperforms the current state-of-the-art (SOTA) HTD methods.**

***Index Terms*—Background-aware learning, data augmentation, hyperspectral target detection (HTD), sparse self-attention mechanism, vision transformer.**

I. Introduction

HYPERSPECTRAL images (HSIs) are acquired by spaceborne or airborne imaging spectrometers, with each pixel containing hundreds of narrow bands that finely capture the spectral reflectance vectors of ground objects [1]. Accordingly, hyperspectral imagery has garnered significant attention in various research areas, including image classification [2], [3], target detection [4], [5], anomaly detection [6], [7], [8], and environmental monitoring [9]. Hyperspectral target detection (HTD) involves using the prior spectral information of the ground object of interest to determine whether each pixel in an HSI belongs to the target by analyzing its spectral information [10]. The reference prior spectra typically come from spectral libraries or identified target pixels within the scene, and usually only a few reference spectra are available [11].

Over the past decades, numerous target detection methods have been developed, which can generally be categorized into three groups: classical methods, sparse representation-based methods, and deep learning-based methods. The classic target detection algorithms typically rely on probabilistic density models, subspace models, or linear mixing models [12], [13], [14]. These algorithms include techniques such as the spectral angle mapper (SAM) [12], matched filter (MF), and constrained energy minimization (CEM) [13]. CEM [13] assumes that HSI has minimal energy, and it introduces a finite impulse filter to enhance the response of the target while suppressing the background signal. However, due to the complexity of the data distribution and the lack of nonlinear representations, the classical detectors often perform poorly in real scenarios [11]. To address these issues, CEM algorithms based on multilevel/multiscale techniques and spatial–spectral feature extraction have been proposed, such as hierarchical CEM (hCEM) [15], ensemble-based cascaded CEM (ECEM) [16], and multiscale superpixel-based CEM [17]. Moreover, Zhao et al. [18] used the fractional Fourier transform to project the spectral into the fractional domain, combining this with a sliding double window strategy to revise the CEM and improve the target detection accuracy. The key idea of the sparse representation-based methods is that each pixel spectrum can be represented as a linear combination of atoms from an overcomplete dictionary [19]. Some researchers have further refined and optimized the sparse representation model from the perspective of collaborative representation [20], [21], background dictionary optimization [22], [24], and tensor decomposition [25], [26]. However, most methods require the construction of both target and background dictionaries, which typically involves more parameter settings and time consumption, significantly reducing the target detection capabilities.

Due to its powerful capability to extract deep discriminative features, deep learning has been successfully and widely applied in HSI processing tasks [27], [28]. More recently, network frameworks such as convolutional neural networks

Received 15 January 2025; revised 13 June 2025 and 12 August 2025; accepted 7 September 2025. Date of publication 10 September 2025; date of current version 19 September 2025. This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project under Grant 2024CSJZN1300, in part by Shanghai Municipal Education Commission Science and Technology Project under Grant 2024AI02002, in part by the National Natural Science Foundation of China under Grant 42171335, and in part by the National Civil Aerospace Project of China under Grant D040102. *(Corresponding author: Kun Tan.)*

Zhiwei Wang and Xue Wang are with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China (e-mail: wangzw@stu.ecnu.edu.cn; wx_ecnu@yeah.net).

Kun Tan is with the Key Laboratory of Geographic Information Science (Ministry of Education) and the School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China (e-mail: tankuncu@gmail.com).

Xiaojun Xu is with Shanghai Environmental Monitoring Center, Shanghai 200003, China (e-mail: sagtsa682@126.com).

Digital Object Identifier 10.1109/TGRS.2025.3608271

(CNNs) [29], [30], autoencoders [31], [32], and transformers [33], [34] have been widely applied in HTD. Concurrently, adversarial learning [35], [36], [37], unsupervised learning, and self-supervised learning [38], [39] approaches are also being applied to learn the complex background distribution. For example, Xie et al. [40] proposed background learning based on target suppression constraint (BLTSC) methods, which fuses adversarial learning and reconstruction error strategy to detect targets. As for the unsupervised methods, Shen et al. [41] proposed a hyperspectral target detection-interpretable representation network (HTD-IRN) that introduces a subspace representation network for estimating endmembers and abundances, combining a transformer module to enhance feature extraction. Self-supervised learning frameworks can capture feature representations of data without relying on labels and can maintain excellent detection performance even during fine-tuned with a small number of samples [42], [43]. For instance, self-supervised spectral-level contrastive learning-based HTD (SCLHTD) [30] and contrastive self-supervised learning-based HTD method with dual path networks (DPN-CSSTD) [44] leverage spectral differences and spatio-spectral correlations in unlabeled data, effectively alleviating the issues of low model transferability and insufficient prior information. However, these methods may still face challenges in fully capturing the intricate spatio-spectral variations inherent in large-scale scenes with highly complex backgrounds.

Since the prior target spectra are usually limited to one or only a few, the scarcity of labeled training samples remains a major challenge for target detection [33], [47]. To address this, data augmentation strategies based on physical models [33], generative adversarial networks (GANs) [45], and diffusion models [46] are used to enlarge training samples, thus enriching sample diversity and optimizing network training. Furthermore, many methods employ Siamese networks and contrastive learning, utilizing the feature similarity between the background and the target pixels for target localization. HTD-Net [29] constructs a two-branch similarity measurement network and is trained with pseudo-labeled samples generated by an improved autoencoder and a linear prediction strategy. The spectral aggregation and separation network (SASN) [48] combines a target band random mask strategy to alleviate the sample imbalance problem, and introduces a triplet-soft loss function to enhance the separation between target and background. Alternatively, transformer networks leverage the self-attention mechanism to capture long-range spectral–spatial dependencies, thereby significantly enhancing feature extraction and representation capability. Li et al. [49] proposed a transformer-inspired stacked GAN network to achieve a more detailed reconstruction of the background. HTD based on transformer via spectral–spatial similarity (HTD-TS^3) [50] and HTDFormer [51] generates pseudo-labeled samples via coarse sample selection and flexible sample augmentation strategies, thereby enabling more effective spectral–spatial feature extraction. Compressive sensing-based triplet transformer detector (CS-TTD) [52] employs a Siamese network with a triplet transformer and uses a combined convolutional network classification module to obtain classification results. Nevertheless, these methods still face limitations in both the diversity of generated samples and their ability to discriminate between background and target feature similarities.

The deep learning-based methods have improved the detection performance from the perspectives of feature discrimination and background learning. However, the current detection methods still have the following drawbacks.

1) The limited availability of target samples cannot satisfy the demand for target instances, leading to an imbalance in the training samples and constraining the effectiveness of the model training. It is therefore crucial to develop a data augmentation strategy that is suitable for rapid target detection in large-scale scenarios.
2) Previous studies have mostly used pixel pairs or pixel blocks for the model training, focusing more on the feature differences between target and background, while neglecting the sparsity and global contextual features of the targets. It is also important to address how to integrate target sparsity characteristics and spatial features to improve the detection accuracy of the model.
3) The existing methods primarily deal with small-area HSIs, but large HSIs have more complex backgrounds, making it difficult to accurately distinguish between target and background. In addition, with large HSIs, the computational cost increases significantly, and the inference speed cannot meet fast processing requirements.

Considering the merits of both the data augmentation and deep learning-based methods, we propose a background-aware sparse transformer network (BASTNet) for HTD. First, the proposed method establishes an image-level training framework for HTD through a spatio-spectral data augmentation approach. The data enhancement strategy enables the model to learn from a more diverse set of samples, thereby improving the precision and robustness of the target detection. We then introduce a Siamese sparse transformer network (S^{2}TNet), which captures more robust target spectral features and the long-range global dependencies by establishing pyramid feature extraction and multiscale feature fusion decoding structures. Due to the inherent sparsity of targets, we introduce a sparse self-attention mechanism to enable more precise target detection by concentrating on the most critical information in the target regions. Moreover, a Siamese network-based differential operation is designed to generate a foreground guidance map that can accurately identify the target's specific location, thus improving the separability of the background and target. Finally, we propose a background-aware learning mechanism, which establishes a composite loss function to more accurately separate the target and background. Specifically, we utilize binary cross-entropy (BCE) loss and dice loss to construct the foreground-guided loss, aiming to enhance segment accuracy and locate the targets.

The main contributions of the BASTNet method are summarized as follows.

1) BASTNet is an image-level end-to-end network that can directly output target detection maps. The method is capable of handling large HSIs and has rapid response capabilities during inference, significantly enhancing the performance in practical applications.

2) A data augmentation strategy based on image patches, which includes spatial random masking and target spectra generation, is proposed to enrich the spatial and spectral diversity of the target samples.
3) An S^2TNet integrated with a background-aware learning mechanism is proposed for HTD. This approach combines foreground and background guidance losses to improve the separability between targets and backgrounds.

The rest of this article is organized as follows. Section II describes related work on data augmentation, target sparsity representation, and background distribution learning. Section III details the basic principles of the proposed BASTNet method. Section IV describes the hyperspectral datasets, experimental analysis, and ablation experiments. Finally, our conclusions are drawn in Section V.

II. Related Work

A. Data Augmentation

A primary challenge in HTD is the limitation of labeled target data, which leads to a severe imbalance between target and background samples during training. To mitigate this issue, various data augmentation techniques have been proposed. For instance, Jiao et al. [33] used a data augmentation method based on the radiative transfer model to constructed sufficient and balanced training samples. Shi et al. [53] proposed a nonlinear spectral synthesis method that simulates the nonlinear mixing of target and background spectra. GANs have also been widely adopted due to their ability to learn complex data distributions and address data imbalance [54]. Gao et al. [45] used a GAN network to generate simulated target and background samples, ensuring stable training of the model. Chen et al. [38] improved spectral different features by spectral masking operations to boost the sensitivity of the model. Zhuang et al. [55] investigated an HTD based on masked autoencoder data augmentation to alleviate the shortage of training data.

Although existing studies have achieved good results in enhancing spectral diversity and alleviating sample scarcity, GAN-based augmentation methods often require additional model training and may result in samples that deviate from the real data distribution. Furthermore, these methods generally overlook the spatial features of targets, which are crucial for accurately distinguishing small targets in large-scale scenes. Although spatial masking strategies exist in the context of anomaly detection [56], [57], such strategies are not directly transferable to supervised target detection due to fundamental differences in the problem formulation. To address this limitation, we proposed a spatio-spectral data augmentation method that can simultaneously improve both the spatial structure and spectral diversity. This approach effectively mitigates the challenges of target sparsity and data imbalance, while avoiding the computational burden and potential instability associated with GAN-based data augmentation methods.

B. Target Sparse Representation

In recent years, sparse representation-based methods have provided innovative approaches for target detection. The main idea is that the spectra can be represented as a linear combinations of atoms from an over-complete dictionary, which can enhance target detection performance [19]. For example, Li et al. [20] combined sparse and collaborative representations and used the residuals of the two representations to detect targets. Cheng et al. [22] proposed a decomposition model with background dictionary learning (DM-BDL), which uses locality-constrained linear coding and comprehensive learning of an a priori target dictionary to suppress the contamination of the background dictionary by target signals. To fully utilize the spatial information in the hyperspectral imagery, Feng et al. [25] proposed a detection algorithm based on low-rank tensor decomposition, which extracts pure background information. Recently, state-of-the-art (SOTA) methods have evolved toward model-driven deep networks, which couple physical priors with deep neural architectures to enhance interpretability. Specifically, the model-driven deep mixture network (MDMN) [58], the low-rank representation network (LRR-Net) [59], and the joint-sparse prior encoding network (JSPEN) [5] transform regularization parameters into trainable parameters within networks to emphasize the interpretability of the model. These approaches achieve competitive performance but still rely heavily on multistage pipelines and manual parameter settings.

Despite these advances, a key technical gap remains: existing SOTA methods generally rely on explicit background dictionary construction, lacking sufficient capability for dynamic parameter adjustment and optimization in complex scenarios. To address these inherent limitations and retaining interpretable sparse prior knowledge, we propose an innovative synergistic integration model that combines sparse representation with a self-attention mechanism. Specifically, the proposed sparse self-attention mechanism dynamically removes background regions by learning sparse mask weights, allowing the network to automatically focus on target features without the need for multistage manual optimization. Specifically, the proposed sparse self-attention mechanism automatically and directly focuses on target features without the need for multistage manual optimization. In addition, we design the low-rank prior as a differentiable loss function, enabling the network to adaptively learn compact low-rank representations of background features while avoiding the high computational overhead of traditional iterative solvers.

C. Background Distribution Learning

Due to the ability of deep learning to automatically learn the nonlinear distribution features of the background, it has an advantage in learning the background distribution of HSIs. The existing advanced methods for background distribution modeling in HSIs mainly involve adversarial learning and contrastive learning paradigms [38]. Adversarial learning leverages the adversarial game between the generator and the discriminator to approximate the true background distribution, potentially improving the model's capability to represent complex backgrounds. For instance, Qin et al. [35] proposed a novel two-stage detection framework based on adversarial learning, which extracts spectral features in latent space through background reconstruction under weak supervision. Zhang et al. [60] integrated the exploiting of variational features in HSIs

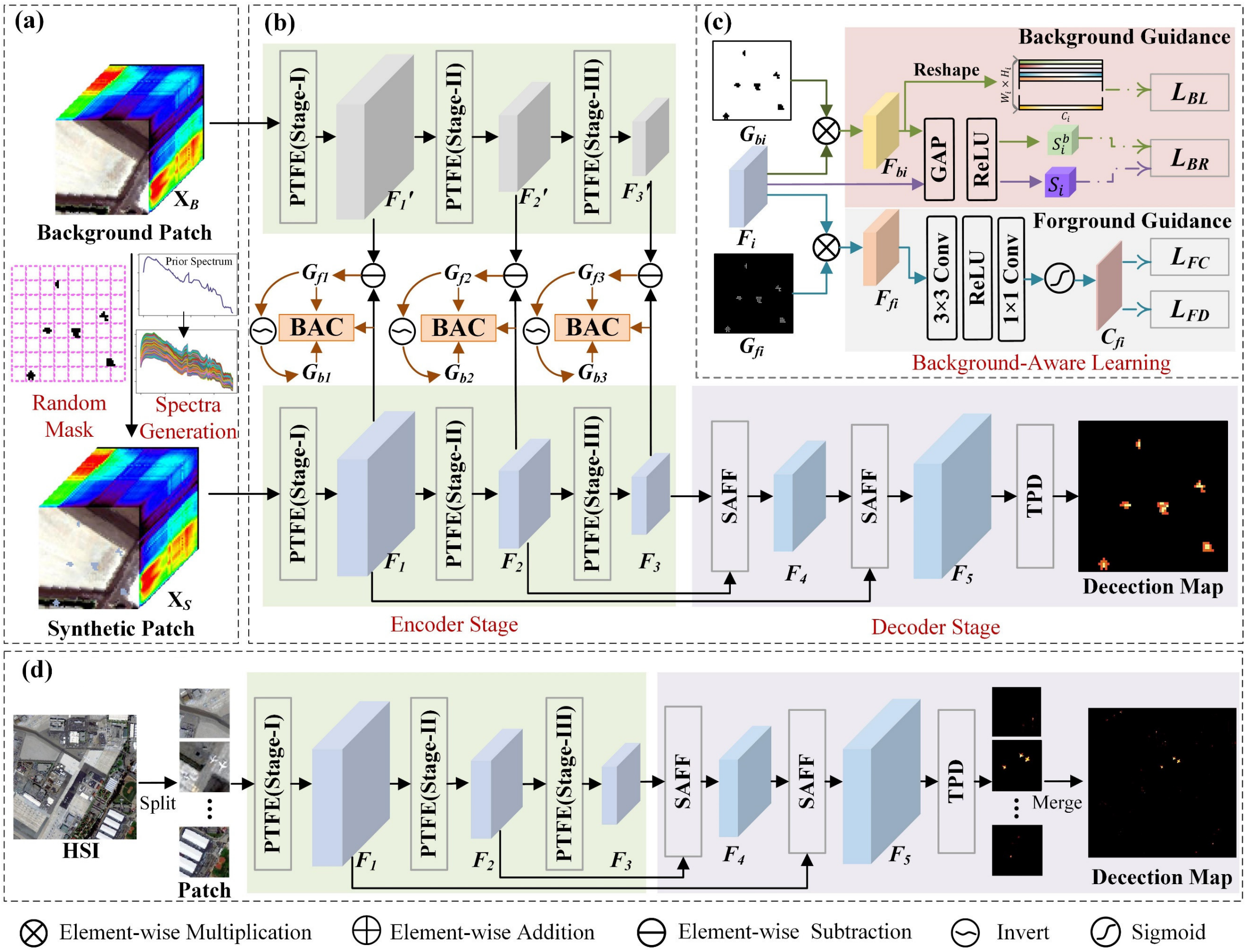


Fig. 1. Overview illustration of the proposed BASTNet method. (a) Training sample generation. (b) S^{2}TNet. (c) BAC. (d) Inference stage.

with supervision of target priors, helping the detector achieve a balance between target detectability and background suppressibility. Despite these advantages, adversarial learning methods still face challenges such as mode collapse and training instability, which can limit their effectiveness and robustness in practical applications. Furthermore, Siamese networks are used to model the separation between background and target, enhancing target detection capability through contrastive loss [44], [45]. For example, Rao et al. [61] employed a Siamese network to solve the problem of similarity metric learning to make homogeneous features as close as possible and heterogeneous features as far as possible. The HTD-Net [29], HTD-TS3 [50], and SASN [48] models learn the feature differences between "target-target" and "target-background" pixel pairs, aiding in the distinction between target and background. However, contrastive learning primarily performs similarity measurement in the spectral feature space, which results in insufficient characterization of spatial features under complex backgrounds.

To accurately learn the background distribution characteristics of large-scale HSIs, we propose an S^{2}TNet to enhance robustness and computational efficiency in complex HSI scenarios. Unlike conventional Siamese networks that rely on explicitly constructed positive–negative sample pairs for similarity metric learning, the Siamese architecture of S^{2}TNet compels the network to automatically separate target and background features in the feature learning phase, without requiring explicit sample similarity metrics. In addition, a background-aware learning mechanism is designed to accurately capture the distribution characteristics of the target and background by imposing background and foreground loss constraints.

III. Proposed Method

A. Overall Framework

The proposed BASTNet method consists of four main components: training sample construction, the S^{2}TNet architecture, the background-aware constraint (BAC) module, and target inference testing. An overview of the architecture of BASTNet is provided in Fig. 1. First, a random masking strategy [56] and a target spectra generation strategy are employed to generate synthetic training samples containing both target and background for each training HSI. The synthetic patches and pure background patches are then fed into S^{2}TNet for training. Finally, during the test phase, the test HSI is split into patches, and each patch is passed through the test network. The patches are then merged to obtain the detection map.

For the training data, both background and target samples are fed into the network in the form of patches for training. The training set consists of two corresponding parts: a pure background HSI patch, denoted as $\boldsymbol{X_B} \in \mathbb{R}^{H\times W\times B}$, and a synthetic patch containing both target and background $\boldsymbol{X_S} \in \mathbb{R}^{H\times W\times B}$, where H, W, and B represent the height, width, and band number of the HSI, respectively. The synthetic patch $\boldsymbol{X_S}$ is generated from the background patch $\boldsymbol{X_B}$ using a random masking strategy combined with a target spectra generation strategy. Both the training and test data are hyperspectral cubes with the same spatial dimensions and spectral bands. Next, we can optimize S^{2}TNet using two types of training sample pairs: $\boldsymbol{X_B}$ and $\boldsymbol{X_S}$.

S^{2}TNet is constructed using a Siamese network with an encoder–decoder structure, which includes a pyramid transformer feature encoder (PTFE) module, a sparse attention feature fusion (SAFF) module, and a target probability detector (TPD) module. The PTFE module is designed to extract features of both the target and background, while the SAFF module and the TPD module are used to generate the target detection map. The BAC module promotes the learning of each subnetwork within S^{2}TNet by utilizing four different loss functions. Specifically, as shown in Fig. 1, the $\boldsymbol{X_B}$ and $\boldsymbol{X_S}$ patches are fed into the PTFE module to extract feature maps at three scales. The two branches of the PTFE module share weights. Subsequently, only the extracted feature corresponding to $\boldsymbol{X_S}$, which contains both target and background, is input into the SAFF module. Through the TPD module, a target prediction map $\boldsymbol{M} \in \mathbb{R}^{H\times W}$ with a 0, 1 distribution can be obtained.

S^{2}TNet utilizes a Siamese architecture integrated with dual-branched PTFE modules to extract features from $\boldsymbol{X_B}$ and $\boldsymbol{X_S}$, ensuring consistent representation of the background feature. Since $\boldsymbol{X_S}$ is synthesized from $\boldsymbol{X_B}$ and shares the same background components with $\boldsymbol{X_S}$, the Siamese architecture drives the network to disregard background perturbation and focus on target-induced feature variations. Moreover, the difference operation of the PTFE module at each stage of feature extraction eliminates the shared background components and preserves the target feature. Consequently, the differential operation generates a foreground guidance map $\mathbf{G}_{\boldsymbol{fi}} \in \mathbb{R}^{H\times W\times 1}$ corresponding to the target distribution, where i denotes the ith stage of the PTFE module. $\mathbf{G}_{\boldsymbol{fi}}$ can be calculated as follows:

$$\begin{aligned} g_{ijk} &= \left\|\boldsymbol{F_{ijk}} - \boldsymbol{F'_{ijk}}\right\|_2 \\ &= \left\|\mathrm{PTFE}_i\left(\boldsymbol{X_{Sjk}}\right) - \mathrm{PTFE}_i\left(\boldsymbol{X_{Bjk}}\right)\right\|_2 \end{aligned} \quad (1)$$

where g_{ijk} represents the guidance value at position (j, k). $\mathbf{G}_{\boldsymbol{fi}}$ is obtained by calculating the l_2 of the feature vector for each pixel. $\boldsymbol{F}$ and $\boldsymbol{F'}$ represent the feature maps obtained from the two branches through the PTFE module. By inverting $\mathbf{G}_{\boldsymbol{fi}}$, the background guidance map $\mathbf{G}_{\boldsymbol{bi}} \in \mathbb{R}^{H\times W\times 1}$ is obtained. Subsequently, $\mathbf{G}_{\boldsymbol{fi}}$, $\mathbf{G}_{\boldsymbol{bi}}$, and the feature map $\boldsymbol{F_i}$ are fed into the BAC module. Finally, the foreground feature map $\boldsymbol{F_{fi}}$ and background feature map $\boldsymbol{F_{bi}}$ can be obtained as follows:

$$\begin{aligned} \boldsymbol{F_{fi}} &= \boldsymbol{F_i} \otimes \mathbf{G}_{\boldsymbol{fi}} \\ \boldsymbol{F_{bi}} &= \boldsymbol{F_i} \otimes \mathbf{G}_{\boldsymbol{bi}} = \boldsymbol{F_i} \otimes (1 - \mathbf{G}_{\boldsymbol{fi}}) \end{aligned} \quad (2)$$

where $\otimes$ represents elementwise multiplication. For $\boldsymbol{F_{fi}}$, the foreground guidance loss, which includes the foreground classification loss L_{FC} and the dice loss L_{FD}, enhances the activation values of the target features and distinguishes the target regions. On the other hand, the background guidance loss, composed of the background low-rank loss L_{BL} and the ratio loss L_{BR}, enables the learning of the distribution characteristics of the background. Therefore, by combining the differential operations of the Siamese network with background and foreground guidance losses, S^{2}TNet effectively differentiates and enhances target features while suppressing irrelevant background information.

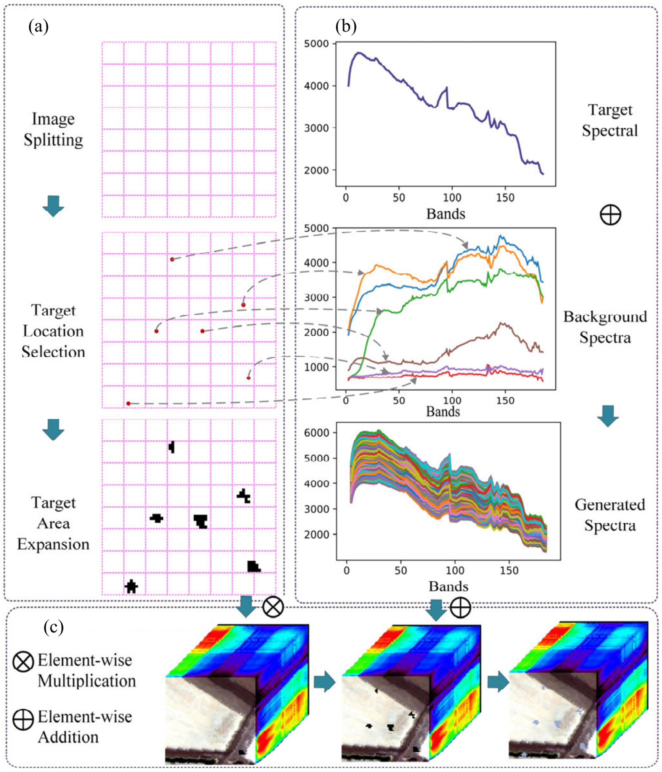


Fig. 2. Flowchart of training sample augmentation. (a) Random masking strategy. (b) Spectra generation strategy. (c) Sample construction.

During the inference phase, the test HSI is first split into 64 × 64 patches. Each patch is then fed into a network composed of the PTFE, SAFF, and TPD modules to generate target probability maps for each patch. Finally, these patches are merged to obtain the final detection map.

B. Training Sample Augmentation

Inspired by the simulation of anomaly samples [57], we designed a data augmentation strategy for HTD that helps BASTNet quickly learn the contextual relationship between target and background. As shown in Fig. 2, the process of training sample augmentation is divided into random mask generation [56] and target spectra generation, which, respectively, simulate the spatial and spectral characteristics of the targets.

1) Random Masking Strategy: As shown in Fig. 2(a), the random masking strategy [56] can simulate the target shape by allowing for irregular shapes and random sizes. The mask map $\boldsymbol{M}$ has the same height H and width W as the input training HSI $\boldsymbol{X_B}$, and $\boldsymbol{M}$ is divided into k^2 nonoverlapping patches, where k is defined as eight. The number of embedded targets N is randomly generated within $[N_{\min}, N_{\max}]$. A pixel position is randomly selected in each selected patches and then the mask map is generated using the iterative expansion method. The target areas are set to one, and the background pixels are set to zero. The region is updated by randomly merging

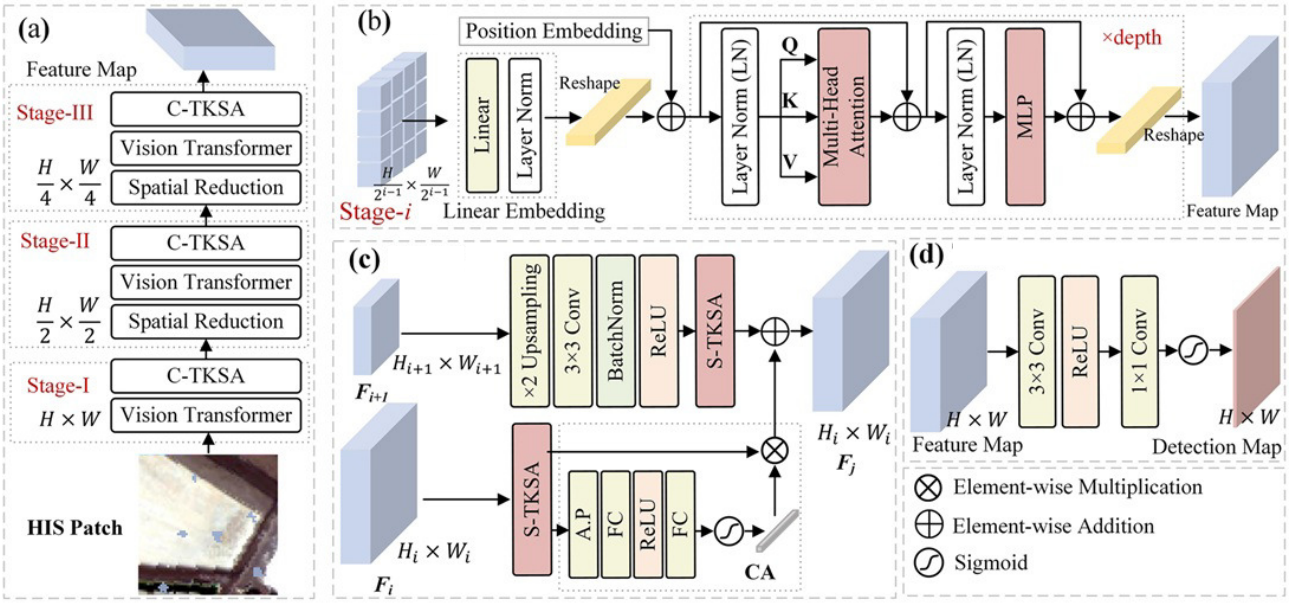


Fig. 3. Illustration of the proposed S^{2}TNet. (a) PTFE. (b) ViT. (c) SAFF. (d) TPD.

until the area reaches A, where A is randomly sampled from $[A_{\min}, A_{\max}]$. Finally, the targets are labeled as one and the background as zero, thus obtaining the ground-truth map of the targets.

2) Target Spectra Generation Strategy: Due to the singularity and insufficiency of the prior target spectra in HTD, we employed a background-target spectral mixing strategy [33], [61] to enhance target samples. As shown in Fig. 2(b), the spectral mixing strategy integrates a random portion of background spectra into the target spectra to simulate realistic variations of targets as real mixed pixels. The prior target and background spectra are denoted as $s_t, s_b \in \mathbb{R}^{B\times 1}$, respectively. s_b is extracted from the pixel in the masked region. The generated spectra s_{tg} based on a linear mixed model can be expressed as follows:

$$s_{tg} = (1-\alpha)\times s_t + \alpha \times s_b + \beta \mathrm{N} \tag{3}$$

where α represents the background abundance, which is randomly set between 0 and 0.3 within the pixels. During each iteration of the network training, a random parameter α is generated for each pixel in the masked region. N denotes the noise vector. The parameter β is defined as 0.05. The target spectra generation strategy significantly increases the number of available target samples and represents the variation of the target spectra under different environments and acquisition conditions.

3) Sample Construction: Finally, the synthetic training sample $\boldsymbol{X_S}$ containing the target and background is built by randomly embedding the generated spectra into $\boldsymbol{X_B}$ with random masks. $\boldsymbol{X_S}$ can be attained by

$$\boldsymbol{X_S} = \boldsymbol{X_B} \otimes \boldsymbol{M} + \boldsymbol{S_t} \otimes (1-\boldsymbol{M}) \tag{4}$$

where $\boldsymbol{S_t} = [\mathrm{s}_{tg}^1, \mathrm{s}_{tg}^2, \ldots, \mathrm{s}_{tg}^n]$ denotes the set of generated spectra, and n is the number of target spectra.

C. Siamese Sparse Transformer Network (S^{2}TNet)

In this section, we introduce the roles and steps of each module in S^{2}TNet, as illustrated in Fig. 3. The PTFE module is designed to extract features at multiple scales, providing more robust feature representations of the targets and background. The SAFF module emphasizes the sparse features of the targets and integrates features from different scales. Finally, the TPD module directly generates the target score map without the need for postprocessing steps.

1) Pyramid Transformer Feature Encoder (PTFE): S^{2}TNet establishes the PTFE module based on a vision transformer (ViT) [62], [63] to effectively extract multiscale features during the encoding process of the HSI. As shown in Fig. 3(a), the PTFE module consists of three feature extraction stages, each comprising a ViT block and a downsampling layer. In the standard self-attention mechanism, the input matrix feature map is linearly transformed to generate three matrices: query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. The attention weights are then computed as follows:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

where d_k is the dimension of both the query and key vectors, which influences the normalization of the attention scores. The ViT block, utilizing a self-attention mechanism, emphasizes the most relevant parts of the image, thereby enhancing the model's ability to identify targets within the hyperspectral data. The downsampling layer, which is a 2 × 2 strided convolutional layer, reduces the spatial resolution of the feature map by half while doubling the number of channels.

At stage i, as depicted in Fig. 3(b), the input image is divided into patches of $(H/2^{i-1})\times(W/2^{i-1})$, which are then processed through a linear projection layer and a normalization layer for patch embedding. These patch feature maps are flattened into vectors, combined with positional embeddings, and reshaped into feature maps. After passing through the feed-forward layer, the final feature map is obtained. The

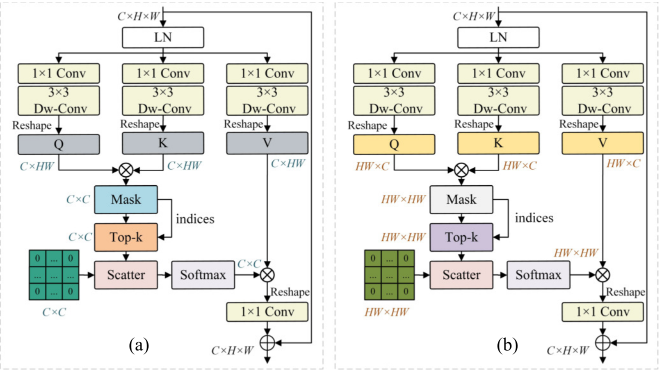


Fig. 4. Top-k sparse attention (TKSA) mechanism. (a) Channel TKSA. (b) Spatial TKSA.

channel numbers in the three feature extraction stages are set to 64, 128, and 512, respectively.

The transformer focuses on global spatial information and captures the long-range dependencies between different parts of the input data. However, it may overlook the channel information in the HSI, failing to fully extract key channel features. Therefore, we introduce the channel top-k sparse attention (C-TKSA) mechanism [64], which aims to more effectively perform the feature aggregation process. Specifically, as shown in Fig. 4(a), 1 ×1 convolutions and 3 × 3 depth-wise convolutions are used to encode the channel-wise context, generating query $\boldsymbol{Q} \in \mathbb{R}^{C\times HW}$, key $\boldsymbol{K} \in \mathbb{R}^{C\times HW}$, and value $\boldsymbol{V} \in \mathbb{R}^{C\times HW}$ representations. The similarity between all the pixel pairs of $\boldsymbol{Q}$ and $\boldsymbol{K}$ is then computed to form the attention matrix $\boldsymbol{M} \in \mathbb{R}^{C\times C}$, which excludes less significant elements with lower weights. Adaptive selection is performed on $\boldsymbol{M}$ to retain the top-k contribution scores, where k is a tunable parameter that dynamically controls the level of sparsity. The top-k sparse selection is as shown in the following equation:

$$A_{ij} = \begin{cases} \dfrac{\text{softmax}\left(\boldsymbol{Q}_i\boldsymbol{K}_j^t\right)}{\sum_{j'\in tk(\boldsymbol{Q}_i\boldsymbol{K}^t)} \text{softmax}\left(\boldsymbol{Q}_i\boldsymbol{K}_{j'}^t\right)}, & \text{if } j \in tk\left(\boldsymbol{Q}_i\boldsymbol{K}^t\right) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where A_{ij} represents the attention score between positions i and j, and $tk(\boldsymbol{Q}_i\boldsymbol{K}^t)$ denotes the set of top-k elements with the highest attention scores. The softmax function is subsequently applied to normalize the k largest elements in each row of the similarity matrix, while the other elements are set to zero. The final sparse self-attention result can be expressed as

$$SA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sum\nolimits_{j\in tk(Q_iK^T)} \boldsymbol{A}_{ij}\boldsymbol{V}_j. \tag{7}$$

The critical parameter of the TKSA mechanism is k, and the optimal selection of k determines the boundary control of the sparsity ratio. A controllable range $[\Delta_1, \Delta_2]$ is established for k, enabling the S^{2}TNet to dynamically adapt and retain the optimal values.

2) Sparse Attention Feature Fusion (SAFF): As shown in Fig. 3(c), the SAFF module is designed to focus more on target activation characteristics, facilitating accurate extraction of target features and their spatial distribution. Operating in a pyramid mode within the decoder, the SAFF module effectively utilizes robust multiscale features and integrates a sparse self-attention mechanism to focus more on the targets themselves. During the upsampling process, the coarser-resolution feature $((H/4) \times (W/4))$ from Stage-III of the decoder and the finer-resolution feature map $((H/2) \times (W/2))$ from Stage-II are input into the SAFF module. This generates a fused feature map with a resolution of $(H/2) \times (W/2)$. Subsequently, the process continues with this fused feature map and the feature map from Stage-I of the encoder, eventually producing the final feature map with resolution $H \times W$.

In each SAFF module, as shown in Fig. 3(c), the coarse-scale feature $\boldsymbol{F_{i+1}}$ is first upsampled using a bilinear interpolation layer, followed by convolution, batch normalization, and rectified linear unit (ReLU) operations, defined as the upsampling processing (UP) block, to obtain a fine-scale feature map. The spatial top-k sparse attention (S-TKSA) mechanism is applied to extract target-related information from the two branches. In addition, in the $\boldsymbol{F_i}$ branch, a channel attention (CA) block [62] is used, which calculates the channel weight vectors to highlight the channels in the feature map. The CA block consists of average pooling, two fully connected layers, a ReLU, and a sigmoid function. Finally, the feature maps from both scales are combined to produce a robust feature map for target detection. The fusion step is shown in the following equation:

$$\begin{aligned} \boldsymbol{F_j} &= TKSA\left(UP\left(\boldsymbol{F_{i+1}}\right)\right) \\ &\quad + TKSA\left(\boldsymbol{F_i}\right) \otimes CA\left(TKSA\left(\boldsymbol{F_i}\right)\right). \end{aligned} \tag{8}$$

In HTD, it is generally assumed that the targets exhibit sparsity [65]. We designed the S-TKSA mechanism to enhance the discriminative ability between target and background. In the S-TKSA mechanism, the attention feature of each pixel is determined by the k most similar pixels, which allows the model to focus more on the targets. As shown in Fig. 4(b), following the initial steps of the S-TKSA mechanism, the query $\boldsymbol{Q} \in \mathbb{R}^{HW\times C}$, key $\boldsymbol{K} \in \mathbb{R}^{HW\times C}$, and value $\boldsymbol{V} \in \mathbb{R}^{HW\times C}$ are generated through two convolution operations. The similarity attention matrix $\boldsymbol{M} \in \mathbb{R}^{HW\times HW}$ is computed between all the pixel pairs. Next, adaptive selection is performed on $\boldsymbol{M}$ to retain the top-k contribution scores, aiming to preserve the most critical target parts while eliminating irrelevant background elements. Finally, the output is obtained by applying the softmax function and matrix multiplication.

3) Target Probability Detector (TPD): The role of the TPD module is to decode the features obtained from the SAFF module into the final target detection map, which represents the target probability of each pixel within each patch. Initially, convolution, batch normalization, and ReLU layers are performed on the feature map to further reduce the number of feature channels. Following this, through convolution operations, the feature channels are reduced to a single channel, and the sigmoid function is applied to generate the final target probability map. In addition, the foreground classifier structure of the BAC module is consistent with that of the TPD module, and the output probability map is utilized to assist in calculating the foreground guidance loss.

D. BASTNet Loss Function

1) Classification Loss: In the training sample construction process, the random masking strategy generates background and target labels. Therefore, the probability of the target and background is calculated using the detection map output from the TPD module. The BCE loss is introduced to supervise the network training process. The loss function is defined as follows:

$$L_{\mathrm{BCE}} = -\frac{1}{\mathrm{HW}}\sum_{i=1}^{\mathrm{HW}}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{9}$$

where y_i denotes the ground-truth label, p_i represents the predicted probability for the target, and HW is the number of samples.

2) Background Guidance Loss: The BAC module is designed with a background guidance loss L_{BG} to capture the distributional characteristics of the hyperspectral background. The L_{BG} loss leverages the low-rank characteristics of the hyperspectral background and is formulated based on the background ratio. In the low-rank representation model, it is assumed that the HSI background exhibits a low-rank property and is typically represented by the nuclear norm of the hyperspectral matrix [65]. The low-rank matrix factorization can be defined as follows:

$$\begin{aligned}&\min_{\boldsymbol{B},\boldsymbol{E}}\ \mathrm{rank}(\boldsymbol{B}) + \lambda\|\boldsymbol{E}\|_1 \Rightarrow \|\boldsymbol{B}\|_* + \lambda\|\boldsymbol{E}\|_1\\ &\text{s.t. } \boldsymbol{X} = \boldsymbol{B} + \boldsymbol{E}\end{aligned} \tag{10}$$

where $\boldsymbol{B}$ represents the background matrix, and rank($\boldsymbol{B}$) is the low-rank constraint function for $\boldsymbol{B}$. $\boldsymbol{E}$ is the sparse matrix, and λ is the balancing parameter. A novel background low-rank loss L_{BL} is proposed to automatically learn low-rank representations of the HSI background by imposing nuclear norm constraints on all feature maps. L_{BL} is defined as follows:

$$L_{\mathrm{BL}} = \frac{1}{L}\sum_{i=1}^{L}\frac{\left\|\boldsymbol{F}_{\boldsymbol{bi}}^{'}\right\|_*}{H_i W_i} \tag{11}$$

where L is the number of feature extraction layers in BASTNet. The feature map $\boldsymbol{F_{bi}}' \in \mathbb{R}^{C\times \mathrm{HW}}$ is transformed by $\boldsymbol{F_{bi}} \in \mathbb{R}^{C\times H\times W}$. $\|\cdot\|_*$ denotes the nuclear norm of the matrix. H_i and W_i are the height and width of the ith feature map $\boldsymbol{F_{bi}}$, respectively. Moreover, the L_{BL} loss normalizes the nuclear norm in each feature map, which helps mitigate the effects of scale variations.

Inspired by background activation suppression [66], we define the background ratio loss L_{BR} as a measure of the discrepancy between the background activation value and the overall feature activation value in terms of their ratio. The L_{BR} loss is used to reduce the influence of background activation values, thereby enhancing the ability of BASTNet to focus on target information. L_{BR} is defined as follows:

$$L_{\mathrm{BR}} = \frac{1}{L}\sum_{i=1}^{L}\frac{S_i^b}{S_i + \varepsilon} = \frac{1}{L}\sum_{i=1}^{L}\frac{\sigma(\mathrm{GAP}(\boldsymbol{F_{bi}}))}{\sigma(\mathrm{GAP}(\boldsymbol{F_i}))+\varepsilon} \tag{12}$$

where ε is set to 10^{-8} to ensure that the equation is well defined. S_i represents the activation value generated from $\boldsymbol{F_i}$. S^b denotes the activation value of the background $\boldsymbol{F_{bi}}$. σ is the ReLU activation function. GAP stands for global average pooling. In summary, as a key part of the BAC module, the L_{BG} loss is achieved by minimizing the predicted probability for background regions.

3) Foreground Guidance Loss: The foreground guidance loss L_{FG} is designed to enhance the model's ability to precisely identify and localize the targets within an image. The L_{FG} loss is composed of two key components: foreground classification loss L_{FC} and dice Loss L_{FD}. L_{FC} provides a robust classification performance by penalizing incorrect predictions, while L_{FD} enhances S^{2}TNet's ability to accurately segment and localize the targets, especially in cases of class imbalance. Since small targets are invisible in Stage-III features, we compute L_{FG} only for the first two stages. By multiplying $\mathbf{G}_{\boldsymbol{fi}}$ with $\boldsymbol{F_i}$, the foreground feature is fed into the foreground classifier of the BAC module. The BCE loss is then calculated from the classification result to obtain L_{FC}. L_{FC} can be denoted as follows:

$$\begin{aligned}L_{\mathrm{FC}} = -\frac{1}{L}\sum_{i=1}^{L}\frac{1}{H_i W_i}\sum_{j=1}^{H_i W_i}&\left[y_j \log(p_j)\right.\\ &\left.+(1-y_j)\log(1-p_j)\right]\end{aligned} \tag{13}$$

where p_j represents the predicted probability of the targets in the foreground classification map. The ground-truth maps at different stages are obtained by downsampling the original ground-truth map.

Dice loss [67] is designed to handle imbalanced samples where the foreground class is relatively small compared to the background. L_{FD} can be formulated as

$$L_{\mathrm{FD}} = \frac{1}{L}\sum_{i=1}^{L} 1 - \frac{2\,|S_i \cap Y|}{|S_i| + |Y|} \tag{14}$$

where $|S_i \cap Y|$ represents the area of overlap between the predicted segmentation S_i and the ground truth Y. $|S_i|$ and $|Y|$ are the areas of the predicted and ground-truth segmentations, respectively. Dice loss emphasizes the overlap between the predicted regions and the actual target regions, ensuring that the model gives sufficient attention to the small target areas. In addition, HSIs can contain noise and irrelevant information, and the dice loss reduces the likelihood of the model misclassifying background areas as target regions.

4) Total Loss: The total loss of BASTNet can be expressed as follows:

$$L = L_{\mathrm{BCE}} + L_{\mathrm{BL}} + L_{\mathrm{BR}} + L_{\mathrm{FC}} + L_{\mathrm{FD}}. \tag{15}$$

By jointly optimizing these three losses, BASTNet effectively integrates information from the target classification, foreground guidance, and background suppression. This approach directs the target prediction map to the target regions, resulting in more precise target detection.

IV. Experiments Results and Analysis

A. Datasets

In the experiments conducted in this study, one public hyperspectral dataset and four large-scene hyperspectral datasets developed in this study were used to validate the proposed BASTNet method.

1) San Diego Dataset: This dataset is publicly available hyperspectral data, acquired by the airborne visible infrared imaging spectrometer (AVIRIS) over the San Diego airport area in California, USA. The San Diego dataset covers a

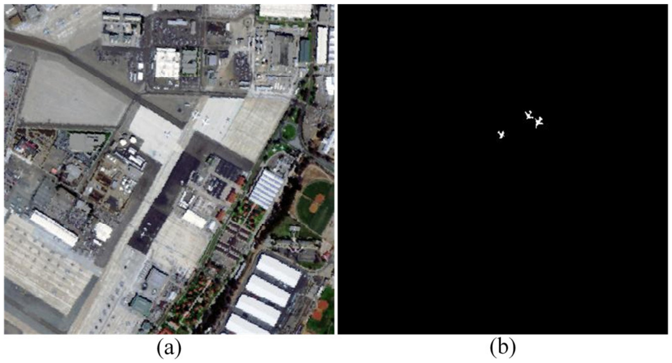


Fig. 5. Pseudo-color images and ground-truth maps for the SanDiego dataset. (a) Pseudo-color image. (b) Ground-truth map.

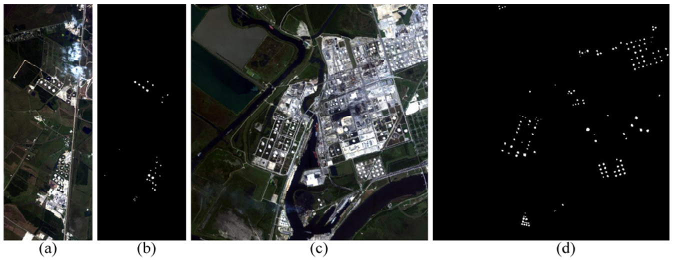


Fig. 6. Pseudo-color images and ground-truth maps for the two Beaumont datasets. (a) Beaumont-I pseudo-color image. (b) Beaumont-II pseudo-color image. (c) Beaumont-I ground-truth map. (d) Beaumont-II ground-truth map.

wavelength range of 370–2510 nm. After removing 35 bands affected by water vapor absorption and a low signal-to-noise ratio, 189 bands were retained [68]. The spatial resolution of the images is 3.5 m. The scene used in the experiment covers an area of 400 × 400 pixels. Fig. 5(a) shows the scene of this dataset. Three airplanes were considered as the targets to be detected, as illustrated in Fig. 5(b).

2) Beaumont Dataset: This dataset was acquired by the airborne visible/infrared imaging spectrometer-next generation (AVIRIS-NG) over the Beaumont area of California, USA, and was downloaded from the AVIRIS-NG website[1] (id: ang20191004t185054rfl). After removing the noisy bands 16–109, 119–145, 159–187, 228–274, and 329–407), 276 bands were retained. The spatial resolution of the dataset is 8.4 m. By cropping the original images, large-scale HSIs of 300 × 800 and 600 × 600 pixels were generated, as shown in Fig. 6(a) and (b). The materials on the top of the storage tanks were considered as the targets to be detected, as shown in Fig. 6(c) and (d).

3) Qingpu Dataset: This dataset was acquired over the Qingpu, Shanghai region of China using the airborne multimodality imaging spectrometer (AMMIS) developed by the Shanghai Institute of Technical Physics at the Chinese Academy of Sciences. The wavelength range is 400–1000 nm, and 250 bands were retained after removing the water vapor and noisy bands. The spatial resolution of the images is 0.75 m. After cropping the original strip data, two HSIs of 400 × 740 pixels were generated, as shown in Fig. 7(a) and (c). The blue rooftops were considered as the targets to be detected, as shown in Fig. 7(b) and (d).

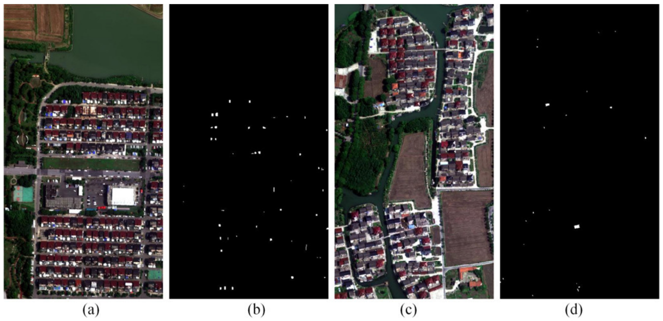


Fig. 7. Pseudo-color images and ground-truth maps for the two AMMIS datasets. (a) Qingpu-I pseudo-color image. (b) Qingpu-I ground-truth map. (c) Qingpu-II pseudo-color image. (d) Qingpu-II ground-truth map.

B. Experimental Setup

1) Comparison Methods: The baseline was made up of seven popular methods of target detection: SAM [12], CEM [13], hCEM[2] [15], DM-BDL[3] [22], BLTSC[4] [40], HTD-IRN[5] [41], and TSTTD[6] [33]. For the hCEM algorithm, parameter λ was set to 200 and the tolerance ε was set to 10^{-6} for all five datasets. For the DM-BDL algorithm, the regularization parameter γ, the number of background atoms m, and the tradeoff parameters λ and β were set to (50, 10, 0.01, 0.1). For the BLTSC algorithm, the positive parameter λ was set to ten, the learning rate was set to 0.0001, and the epochs were set to 500. For the HTD-TRN algorithm, the background subspace dimension m and the sparsity weight η_1 were set to (5, 0.001) for all five datasets. For the TSTTD algorithm, the batch size was 64, the learning rate was 0.0001, and the weight decay was zero.

2) Experimental Settings: The hardware device used in the experiments was a computer with an Intel Core i7-10700K CPU @ 3.80 GHz, an NVIDIA GeForce RTX 3070 GPU, and 64-GB RAM. For the software environment, the hCEM and DM-BDL detection methods were implemented using MATLAB R2020b, while Python 3.9 was used for the SAM, CEM, BLTSC, HTD-IRN, and TSTTD methods. The deep learning algorithms were implemented using Torch-GPU 2.0.1 and CUDA 11.3. We utilized the Adam optimizer for BASTNet, and set the learning rate to 0.001 and the weight decay to 0.0002.

For the San Diego dataset, we selected regions of interest within the background area and applied the image operations of cropping, rotation, and flipping, thereby expanding the training samples to 305. For the Beaumont and Qingpu datasets, operations such as cropping were applied to the background areas of the original strips, resulting in expanded training samples of 565 and 484, respectively. The training sample size for BASTNet was fixed at 64 × 64, with the number of pseudo-targets ranging from $[N_{\min}, N_{\max}]$ set to [3] and [5], and the area range $[A_{\min}, A_{\max}]$ set to [3] and [10]. For C- TKSA, $[\Delta_1, \Delta_2]$ was set to $[(1/2), (4/5)]$. For S-TKSA, $[\Delta_1, \Delta_2]$ was also set to $[(1/100), (1/10)]$, with the value determined by the ratio of the number of targets to the number of pixels.

[1] https://avirisng.jpl.nasa.gov/

[2] https://levir.buaa.edu.cn/Code.htm

[3] https://github.com/FDU-ctk/HSI-detection

[4] https://github.com/zhangxin-xd/BLTSC/tree/master

[5] https://github.com/shendb2022/HTD-IRN

[6] https://github.com/shendb2022/TSTTD

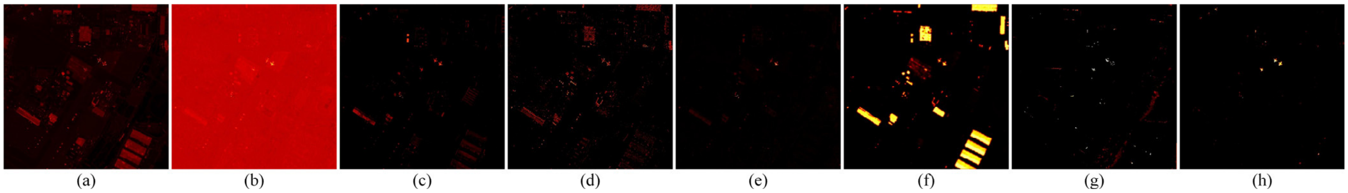


Fig. 8. Detection maps for the San Diego dataset. (a) SAM. (b) CEM. (c) hCEM. (d) DM-BDL. (e) BLTSC (f) HTD-IRN. (g) TSTTD. (h) BASTNet.

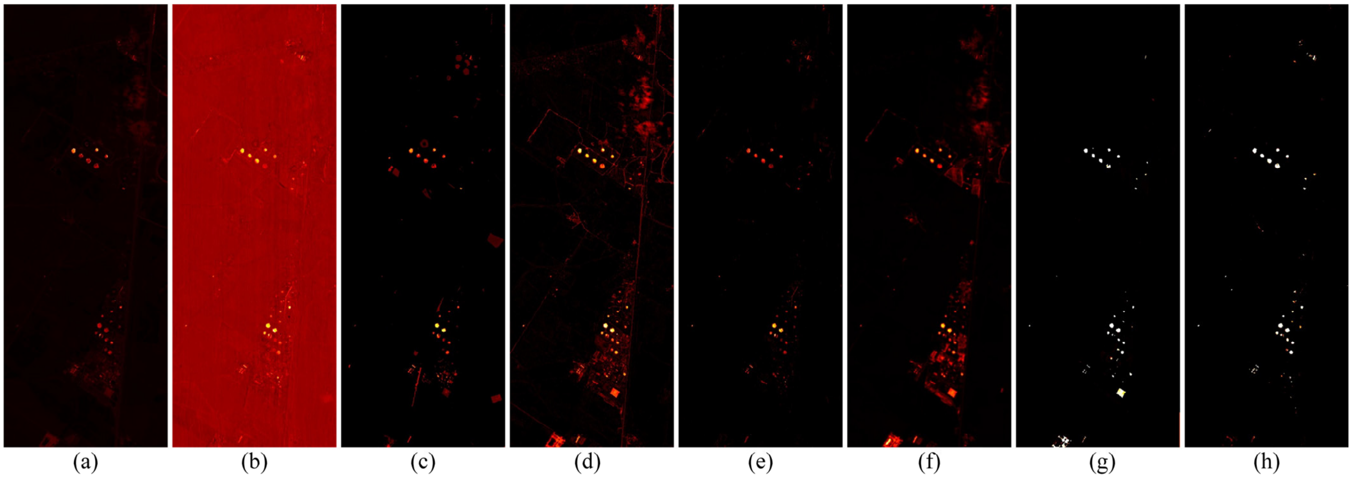


Fig. 9. Detection maps for the Beaumont-I dataset. (a) SAM (b) CEM. (c) hCEM. (d) DM-BDL. (e) BLTSC. (f) HTD-IRN. (g) TSTTD. (h) BASTNet.

3) Evaluation Metrics: To comprehensively evaluate the performance of the target detection algorithms, 3-D receiver operating characteristic (ROC) curves [69], the area under the curve (AUC) scores [69], and box-whisker plots [33] are used in this article. The 3-D ROC curves were plotted using the detection probability P_d, the false alarm rate P_f, and the threshold τ. From the 3-D ROC curves, three types of 2-D ROC curves can be derived: (P_d, P_f), (P_d, τ), and (P_f, τ), which, respectively, evaluate the algorithm's detection efficiency, background suppression, and overall performance. To quantitatively evaluate the performance of the algorithms, seven AUC values are used here for the assessment: $\text{AUC}_{(P_d,P_f)}$, $\text{AUC}_{(P_d,\tau)}$, $\text{AUC}_{(P_f,\tau)}$, AUC_{TD}, AUC_{BS}, AUC_{ODP}, AUC_{TDBS}, and AUC_{SNPR}. $\text{AUC}_{(P_d,P_f)}$, $\text{AUC}_{(P_d,\tau)}$, and $\text{AUC}_{(P_f,\tau)}$ were calculated based on the AUC of the three 2-D ROC curves, while the remaining AUC values were derived from $\text{AUC}_{(P_d,P_f)}$, $\text{AUC}_{(P_d,\tau)}$, and $\text{AUC}_{(P_f,\tau)}$. The expressions are as follows:

$$\text{AUC}_{\text{TD}} = \text{AUC}_{(\text{D,F})} + \text{AUC}_{(\text{D},\tau)} \tag{16}$$

$$\text{AUC}_{\text{BS}} = \text{AUC}_{(P_d,P_f)} - \text{AUC}_{(P_f,\tau)} \tag{17}$$

$$\text{AUC}_{\text{ODP}} = \text{AUC}_{(P_d,P_f)} + \text{AUC}_{(P_d,\tau)} - \text{AUC}_{(P_f,\tau)} \tag{18}$$

$$\text{AUC}_{\text{TDBS}} = \text{AUC}_{(P_d,\tau)} - \text{AUC}_{(P_f,\tau)} \tag{19}$$

$$\text{AUC}_{\text{SNPR}} = \frac{\text{AUC}_{(P_d,\tau)}}{\text{AUC}_{(P_f,\tau)}} \tag{20}$$

where $\text{AUC}_{(P_d,P_f)}$, $\text{AUC}_{(P_d,\tau)}$, and AUC_{TD} are used to evaluate the target detection capability, where higher values indicate a better performance. For the background suppression performance, $\text{AUC}_{(P_f,\tau)}$ and AUC_{BS} assess the background suppression capability, with smaller values of the former and larger values of the latter indicating a better performance. AUC_{ODP} and AUC_{TDBS} reflects the algorithm's robustness and stability. AUC_{SNPR} reflects the signal-to-noise ratio of the algorithm. Higher values of AUC_{ODP} and AUC_{SNPR} indicate a better overall performance and target detection capability.

C. Detection Results

In this section, we validate the effectiveness of the proposed BASTNet method and compare its detection performance with that of the other methods. Figs. 8–13 provide a qualitative evaluation of various methods across the five datasets. Tables I–V present the AUC values for the different images.

Figs. 8–11 show a detection map of various methods across the five datasets. It can be observed that all three aircraft targets in the San Diego dataset are detected by all the methods. Both BLTSC and BASTNet demonstrate excellent background suppression capabilities. However, hCEM, DM-BDL, and TSTTD exhibit some false alarms, and TSTTD also shows missing pixels in certain targets. HTD-IRN mistakenly identifies the rooftops of buildings as targets, primarily due to the complex background characteristics. Notably, the proposed BASTNet method not only accurately detects the edge information of the targets but also performs well in suppressing the complex background. For the two Beaumont datasets. SAM, CEM, hCEM, and BLTSC struggle to identify the targets of different scales. In addition, BLTSC and HTD-IRN are less effective at highlighting the targets, with HTD-IRN exhibiting false alarms on the Beaumont-I dataset. TSTTD and BASTNet show outstanding performances, with TSTTD demonstrating the best background suppression, while BASTNet performs the best in target detection. For the two Qingpu datasets, SAM and CEM show the worst detection performance because they do not effectively suppress the background. Although hCEM and DM-BDL show similar detection performances across the two datasets, the background suppression is poor in the

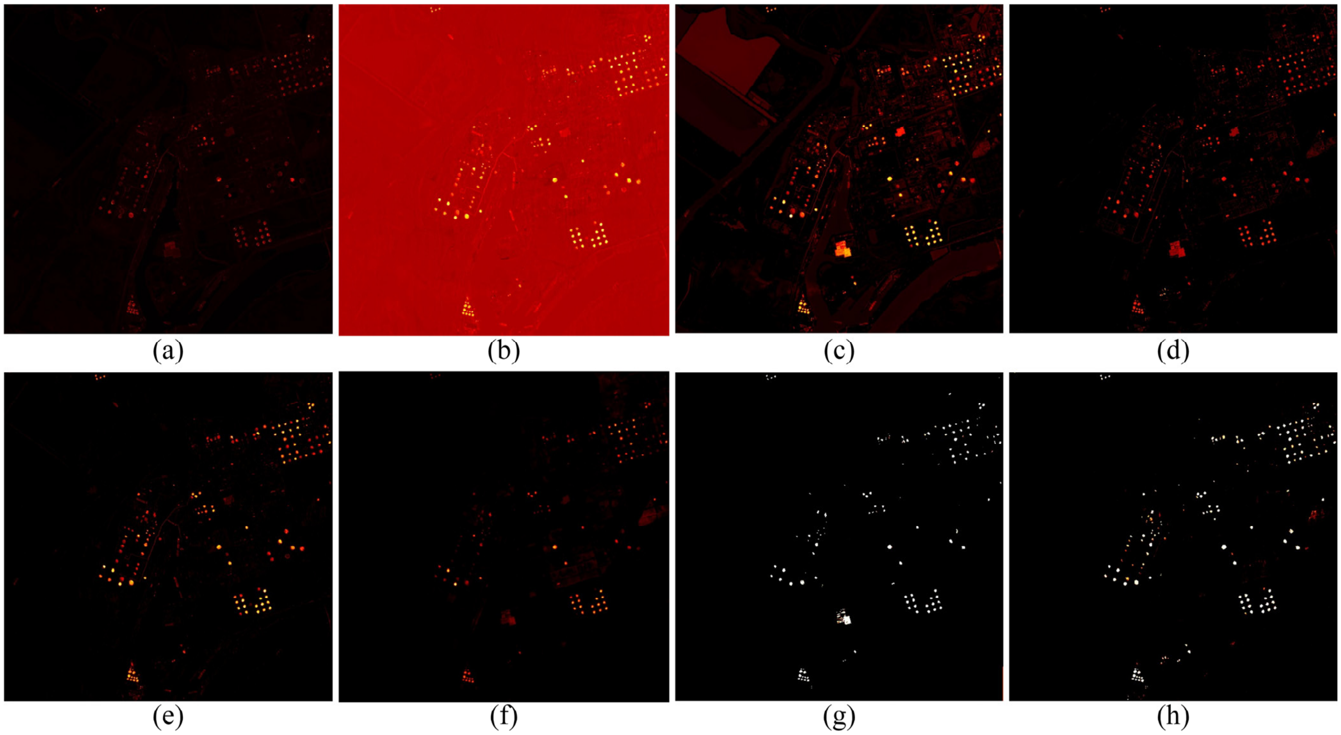


Fig. 10. Detection maps for the two AMMIS datasets, with the first row being Qingpu-I and the second row Qingpu-II. (a) SAM. (b) CEM. (c) hCEM. (d) DM-BDL. (e) BLTSC. (f) HTD-IRN. (g) TSTTD. (h) BASTNet.

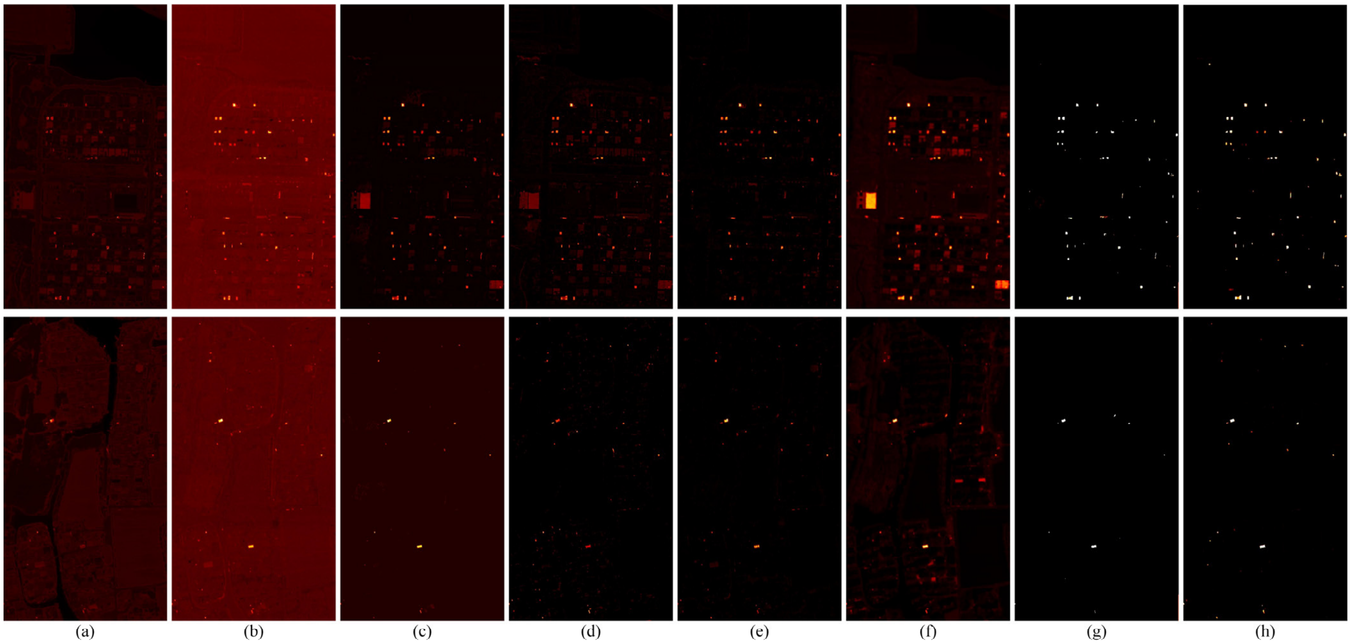
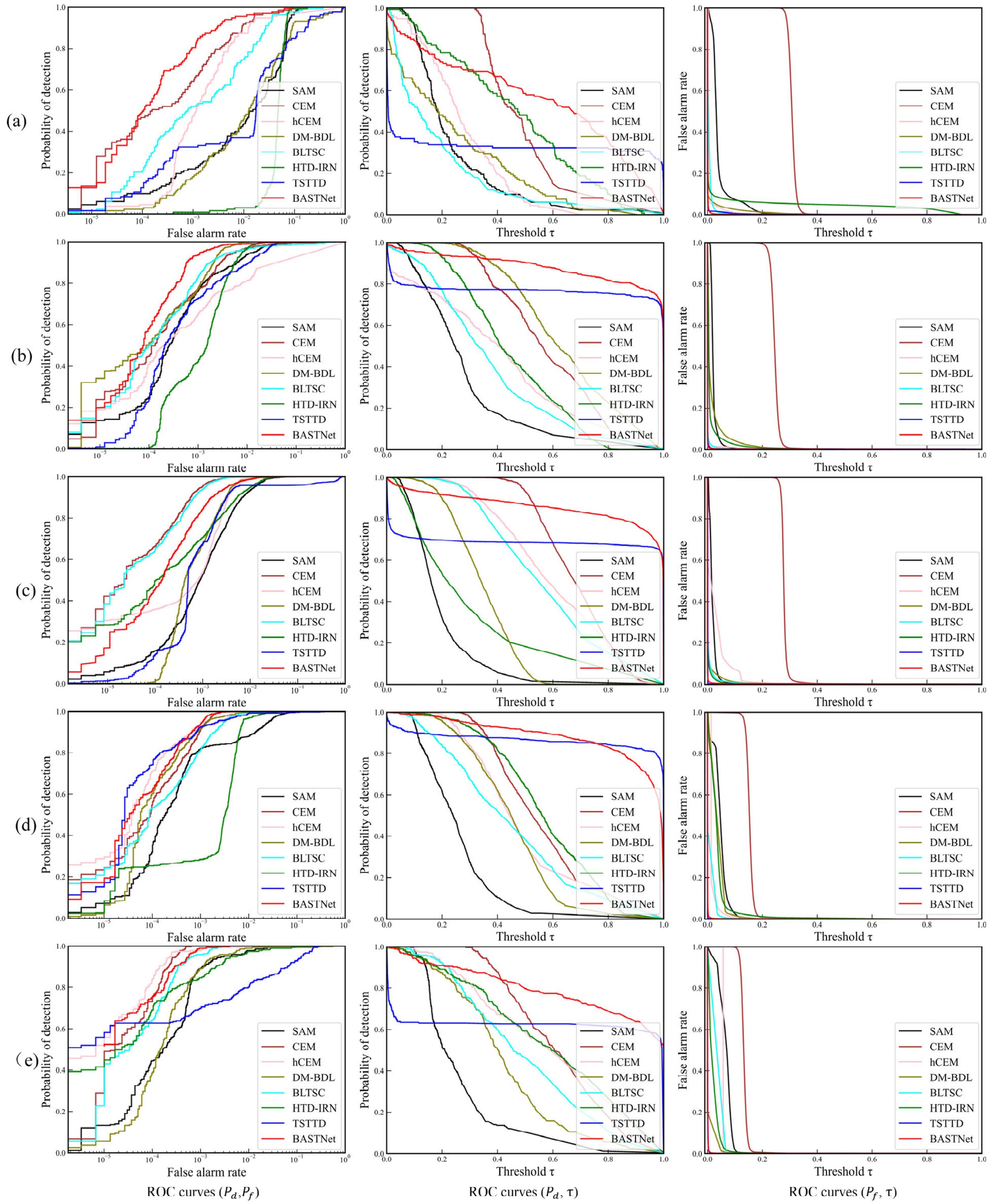


Fig. 11. Detection maps for the Beaumont-II dataset. (a) SAM. (b) CEM. (c) hCEM. (d) DM-BDL. (e) BLTSC. (f) HTD-IRN. (g) TSTTD. (h) BASTNet.

Qingpu-I dataset. BLTSC can effectively localize targets, but it struggles to maintain the accurate shapes of the targets. HTD-IRN can highlight the targets but tends to identify more buildings as targets. TSTTD and BASTNet demonstrate a higher confidence in target detection, but TSTTD has difficulty detecting the small-scale targets in the Qingpu-II dataset. Overall, the proposed BASTNet method not only detects the targets of varying scales, but also suppresses the background to a shallow level.

Fig. 12 shows the three types of ROC curves for the five datasets. The ROC curve of (P_d, P_f) is close to the top-left corner, and the ROC curve of (P_d, τ) is close to the top-right corner, which indicates a better detection performance. The ROC curve of (P_f, τ) is near the bottom-left corner, implying that the algorithm has stronger background suppression capabilities. Based on the ROC curves of (P_d, P_f) for the five datasets, the curve of BASTNet is primarily located in the top-left corner, indicating that the algorithm

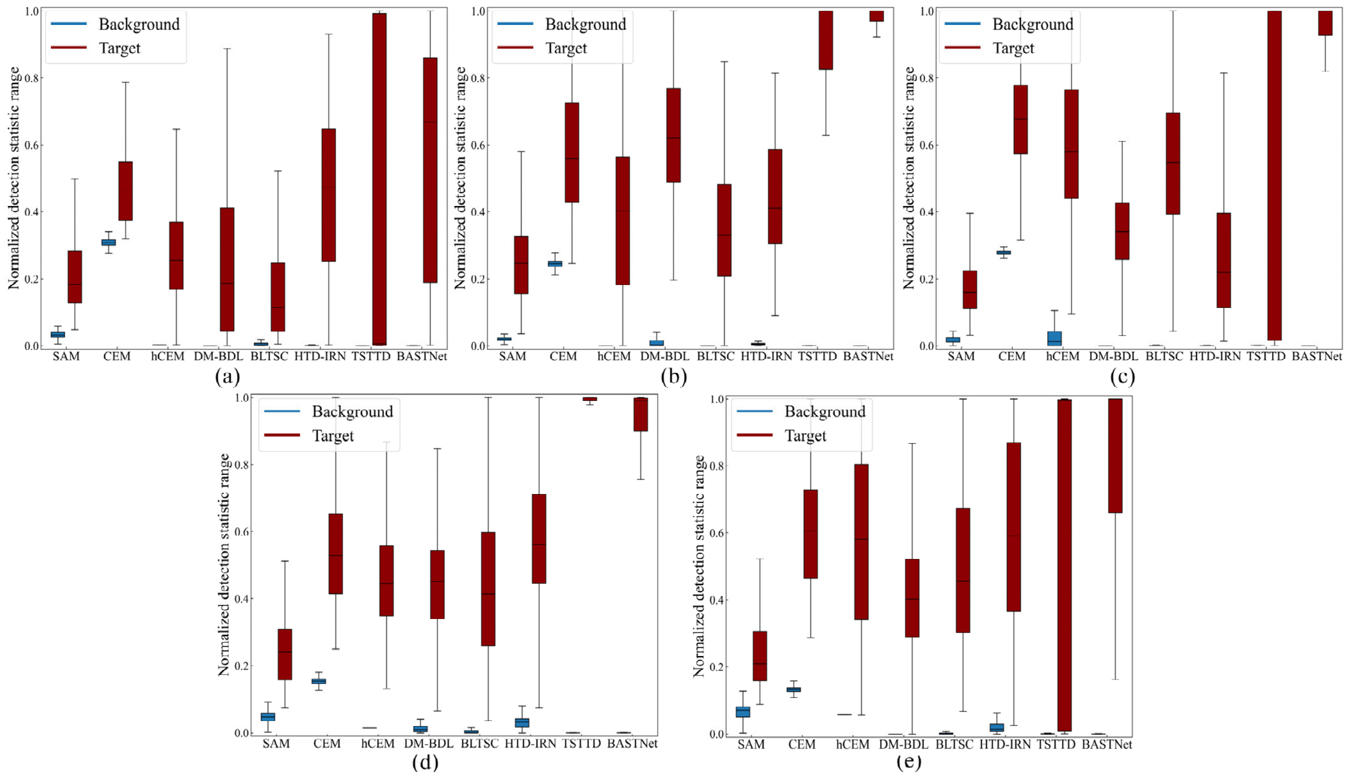


ROC curves (P_d, P_f) ROC curves (P_d, τ) ROC curves (P_f, τ)

Fig. 12. ROC curves for each detection method on the five datasets. (a) San Diego. (b) Beaumont-I. (c) Beaumont-II. (d) Qingpu-I. (e) Qingpu-II.

shows a good detection performance. For the ROC curves of (P_d, τ), BASTNet is generally positioned in the top-right corner, clearly outperforming the other detection algorithms. Even with a high threshold, BASTNet can still maintain a high detection probability. From the ROC curves of (P_f, τ) for the five datasets, the curves of BASTNet and TSTTD are almost overlapping, positioned in the optimal bottom-left corner, which indicates their ability to keep the background at a relatively low level. In summary, the proposed BASTNet not only effectively detects targets but also suppresses the background well, demonstrating the algorithm's excellent robustness.

The box-whisker plots for the five datasets are shown in Fig. 13. The blue and dark red boxes represent the background and target, respectively, with the gap between the two boxes indicating the degree of separation between the target and



(a)

(b)

(c)

(d)

(e)

Fig. 13. Box-whisker plots for each detection method on the five datasets. (a) San Diego. (b) Beaumont-I. (c) Beaumont-II. (d) Qingpu-I. (e) Qingpu-II.

TABLE I

AUC Scores of the Different Methods on the SanDiego Dataset

Method	$\mathrm{AUC}_{(P_d,P_f)}$	$\mathrm{AUC}_{(P_d,\tau)}$	$\mathrm{AUC}_{(P_f,\tau)}$	$\mathrm{AUC}_{\mathrm{TD}}$	$\mathrm{AUC}_{\mathrm{BS}}$	$\mathrm{AUC}_{\mathrm{ODP}}$	$\mathrm{AUC}_{\mathrm{TDBS}}$	$\mathrm{AUC}_{\mathrm{SNPR}}$	Time/s
SAM	0.9703	0.2385	0.0417	1.2088	0.9286	1.1671	0.1968	5.71	1.14
CEM	0.9957	0.4706	0.2897	1.4664	0.7059	1.1766	0.1809	1.62	0.17
hCEM	0.9768	0.2729	0.0069	1.2497	0.9699	1.2428	0.2660	39.48	3.42
DM-BDL	0.9416	0.2408	0.0077	1.1824	0.9339	1.1747	0.2331	31.43	113.77
BLTSC	0.9880	0.1873	0.0058	1.1754	0.9822	1.1695	0.1849	31.98	5.99
HTD-IRN	0.9514	0.4595	0.0453	1.4109	0.9061	1.3656	0.4142	10.15	3.20
TSTTD	0.9386	0.3299	0.0026	1.2686	0.9359	1.2659	0.3273	125.24	52.43
BASTNet	**0.9973**	**0.5600**	**0.0012**	**1.5572**	**0.9960**	**1.5561**	**0.5588**	**451.90**	1.87

TABLE II

AUC Scores of the Different Methods on the Beaumont-I Dataset

Method	$\mathrm{AUC}_{(P_d,P_f)}$	$\mathrm{AUC}_{(P_d,\tau)}$	$\mathrm{AUC}_{(P_f,\tau)}$	$\mathrm{AUC}_{\mathrm{TD}}$	$\mathrm{AUC}_{\mathrm{BS}}$	$\mathrm{AUC}_{\mathrm{ODP}}$	$\mathrm{AUC}_{\mathrm{TDBS}}$	$\mathrm{AUC}_{\mathrm{SNPR}}$	Time/s
SAM	0.9975	0.2822	0.0219	1.2798	0.9756	1.2578	0.2603	12.87	1.81
CEM	0.9965	0.5739	0.2453	1.5705	0.7512	1.3251	0.3286	2.33	0.57
hCEM	0.9324	0.3954	0.0020	1.3278	0.9304	1.3258	0.3993	193.84	6.63
DM-BDL	0.9993	0.6319	0.0175	1.6313	0.9817	1.6137	0.6144	36.01	29.19
BLTSC	0.9943	0.3669	0.0014	1.3613	0.9929	1.3598	0.3655	256.58	7.27
HTD-IRN	0.9976	0.4449	0.0128	1.4425	0.9847	1.4297	0.4320	34.56	3.17
TSTTD	0.9968	0.7679	0.0025	1.7648	0.9943	1.7622	0.7655	311.26	118.83
BASTNet	**0.9994**	**0.8837**	**0.0012**	**1.8831**	**0.9984**	**1.8819**	**0.8827**	**866.41**	1.87

the background. It can be observed that BASTNet achieves the best background suppression while maintaining the maximum separation between the target and the background. The separation between the target and background is relatively small for SAM and CEM, and the background has a broader range. The hCEM, DM-BDL, and HTD-IRN methods show varying degrees of background suppression ability. In the case of TSTTD, the target data include some pixels that are difficult to distinguish from the background, which increases the variability and results in a longer box plot. Overall, BASTNet outperforms the other comparison algorithms and effectively separates the background from the target.

Tables I–V present the AUC scores for each detection method across the five datasets. For the San Diego dataset,

TABLE III

AUC SCORES OF THE DIFFERENT METHODS ON THE BEAUMONT-II DATASET

Method	AUC$_{(P_d,P_f)}$	AUC$_{(P_d,\tau)}$	AUC$_{(P_f,\tau)}$	AUC$_{\text{TD}}$	AUC$_{\text{BS}}$	AUC$_{\text{ODP}}$	AUC$_{\text{TDBS}}$	AUC$_{\text{SNPR}}$	Time/s
SAM	0.9972	0.1891	0.0182	1.1863	0.9791	1.1681	0.1709	10.41	2.61
CEM	**0.9997**	0.6798	0.2798	1.6796	0.7199	1.3997	0.3999	2.42	0.94
hCEM	0.9986	0.5901	0.0280	1.5887	0.9706	1.5607	0.5621	21.09	9.76
DM-BDL	0.9987	0.3412	0.0049	1.3399	0.9938	1.3350	0.3363	69.37	43.25
BLTSC	**0.9997**	0.5609	0.0043	1.5607	0.9954	1.5563	0.5566	129.91	8.92
HTD-IRN	0.9981	0.2995	0.0027	1.2977	0.9955	1.2949	0.2969	113.11	2.09
TSTTD	0.9770	0.6874	0.0024	1.6644	0.9746	1.6620	0.6850	286.00	169.27
BASTNet	0.9990	**0.8597**	**0.0017**	**1.8587**	**0.9973**	**1.8580**	**0.8510**	**506.38**	2.13

TABLE IV

AUC SCORES OF THE DIFFERENT METHODS ON THE QINGPU-I DATASET

Method	AUC$_{(P_d,P_f)}$	AUC$_{(P_d,\tau)}$	AUC$_{(P_f,\tau)}$	AUC$_{\text{TD}}$	AUC$_{\text{BS}}$	AUC$_{\text{ODP}}$	AUC$_{\text{TDBS}}$	AUC$_{\text{SNPR}}$	Time/s
SAM	0.9962	0.2551	0.0459	1.2512	0.9503	1.2054	0.2091	5.55	3.21
CEM	0.9996	0.5538	0.1541	1.5535	0.8456	1.3993	0.3997	3.59	0.63
hCEM	0.9962	0.4886	0.0189	1.4882	0.9808	1.4659	0.4697	25.90	17.47
DM-BDL	0.9996	0.4471	0.0355	1.4467	0.9641	1.4112	0.4116	12.60	86.41
BLTSC	0.9993	0.4406	0.0095	1.4399	0.9898	1.4304	0.4311	239.97	6.87
HTD-IRN	0.9964	0.5658	0.0363	1.5622	0.9600	1.5259	0.5295	15.57	3.75
TSTTD	0.9982	0.8655	**0.0008**	1.8638	0.9974	1.8629	0.8647	**1084.88**	75.91
BASTNet	**0.9998**	**0.8858**	0.0015	**1.8856**	**0.9982**	**1.8841**	**0.8843**	591.19	2.38

TABLE V

AUC SCORES OF THE DIFFERENT METHODS ON THE QINGPU-II DATASET

Method	AUC$_{(P_d,P_f)}$	AUC$_{(P_d,\tau)}$	AUC$_{(P_f,\tau)}$	AUC$_{\text{TD}}$	AUC$_{\text{BS}}$	AUC$_{\text{ODP}}$	AUC$_{\text{TDBS}}$	AUC$_{\text{SNPR}}$	Time/s
SAM	0.9986	0.2673	0.0650	1.2659	0.9336	1.2009	0.2022	4.10	3.21
CEM	0.9998	0.6377	0.1332	1.6376	0.8677	1.5043	0.5045	4.78	0.62
hCEM	0.9943	0.5743	0.0577	1.5686	0.9366	1.5109	0.5166	9.95	5.16
DM-BDL	0.9994	0.4449	0.0088	1.4439	0.9907	1.4355	0.4357	50.36	85.17
BLTSC	0.9997	0.4899	0.0340	1.4897	0.9658	1.4556	0.4559	14.41	6.96
HTD-IRN	0.9974	0.5769	0.0204	1.5743	0.9770	1.5539	0.5566	28.33	4.05
TSTTD	0.9800	0.6233	**0.0008**	1.6033	0.9792	1.6025	0.6225	334.42	76.04
BASTNet	**0.9999**	**0.7996**	0.0009	**1.7995**	**0.9989**	**1.7986**	**0.7898**	**893.87**	2.39

BASTNet obtains the highest AUC scores across all seven metrics, with the AUC$_{\text{ODP}}$ score being 0.2314 higher than that of the second-best TSTTD method. Although CEM shows high AUC$_{(P_d,\tau)}$ and AUC$_{\text{TD}}$ scores, its AUC$_{\text{BS}}$ score is the lowest. On the two Beaumont datasets, BASTNet achieves the highest AUC$_{\text{ODP}}$ scores, reaching 1.8819 and 1.8580, respectively, surpassing the other methods. Although the AUC$_{(P_d,P_f)}$ score for the Beaumont-II dataset is suboptimal, the proposed method still attains optimal values for AUC$_{\text{TD}}$ and AUC$_{\text{BS}}$, demonstrating the comprehensive performance of BASTNet. In the two Qingpu datasets, BASTNet achieves the highest AUC scores in AUC$_{(P_d,P_f)}$, AUC$_{(P_d,\tau)}$, AUC$_{\text{TD}}$, AUC$_{\text{BS}}$, and AUC$_{\text{ODP}}$, demonstrating its excellent overall performance. TSTTD also demonstrates strong capabilities in both target detection and background suppression. In summary, BASTNet confidently detects targets and effectively suppresses the background, achieving precise separation between targets and background.

Moreover, Tables I–V list the inference times of the detection algorithms. Among the traditional algorithms, SAM, CEM, and hCEM maintain relatively fast computation times but exhibit instability, particularly in complex background scenarios. The DM-BDL algorithm has a longer computation time, but its detection performance is superior to those of the traditional algorithms. Among the four deep learning methods, BLTSC involves two processing steps during the test phase, which complicates direct target detection. Although the TSTTD method demonstrates an excellent detection performance, it has the longest inference time. In contrast, BASTNet achieves the fastest inference time while maintaining a high accuracy, thanks to its patch-based approach for rapid inference. Overall, BASTNet maintains a reasonable computation time while delivering an outstanding detection performance, which makes it highly promising for practical applications.

D. Ablation Study

1) Class Activation Map Analysis: The class activation map (CAM) helps explain the decision-making process of the network by showing which regions of the image the network relied on to make the classification decision. As shown in Fig. 14, we present the CAMs for the different phases of

TABLE VI

$\text{AUC}_{(P_d,P_f)}$ AND $\text{AUC}_{(P_f,\tau)}$ SCORES FOR THE DIFFERENT MODULES ON THE FIVE DATASETS

Methods	San Diego	Beaumont-I	Beaumont-II	Qingpu-I	Qingpu-II
Baseline	0.9869/0.0051	0.9919/0.0043	0.9967/0.0031	0.9921/0.0047	0.9990/0.0044
Baseline+BAC	0.9937/0.0013	0.9963/**0.0010**	0.9951/**0.0007**	0.9975/**0.0015**	0.9992/0.0029
Baseline+TKSA	0.9899/0.0013	0.9913/0.0016	0.9967/0.0031	0.9995/0.0037	0.9990/0.0024
Baseline+BAC+TKSA	**0.9973**/**0.0012**	**0.9994**/0.0012	**0.9990**/0.0017	**0.9998**/**0.0015**	**0.9999**/**0.0009**

TABLE VII

$\text{AUC}_{(P_d,P_f)}$ AND $\text{AUC}_{(P_f,\tau)}$ SCORES EACH LOSS TERM ON THE FIVE DATASETS

Loss term	San Diego	Beaumont-I	Beaumont-II	Qingpu-I	Qingpu-II
L_{BCE}	0.9899/0.0013	0.9913/0.0016	0.9967/0.0031	0.9995/0.0037	0.9990/0.0024
$L_{BCE} + L_{BL}$	0.9908/0.0015	0.9980/0.0014	0.9952/0.0025	0.9995/0.0015	0.9997/0.0015
$L_{BCE} + L_{BR}$	0.9906/**0.0011**	0.9989/0.0018	0.9944/0.0026	**0.9998**/0.0016	0.9998/0.0010
$L_{BCE} + L_{BG}$	0.9912/0.0014	0.9988/0.0017	0.9974/0.0021	**0.9998**/**0.0011**	0.9998/**0.0007**
$L_{BCE} + L_{FC}$	0.9955/0.0015	0.9992/0.0018	**0.9991**/0.0023	0.9996/0.0017	0.9995/0.0018
$L_{BCE} + L_{FD}$	0.9927/0.0018	0.9985/**0.0012**	0.9987/0.0030	0.9993/0.0020	0.9998/0.0018
$L_{BCE} + L_{FG}$	0.9962/0.0015	0.9993/0.0015	0.9988/0.0030	0.9996/0.0015	0.9998/0.0010
$L_{BCE} + L_{BG} + L_{FG}$	**0.9973**/0.0012	**0.9994**/**0.0012**	0.9990/**0.0017**	**0.9998**/0.0015	**0.9999**/0.0009

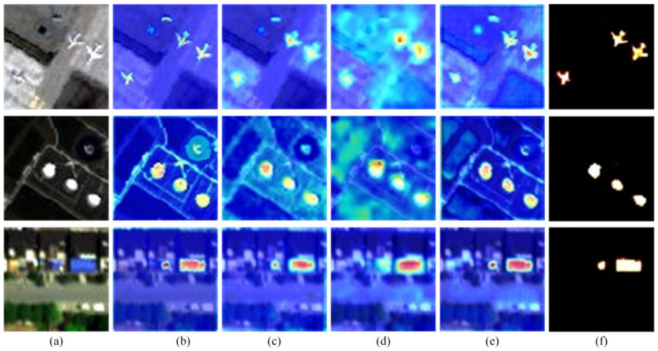


Fig. 14. Comparison of the CAMs from the different stages of S^2TNet. (a) Pseudo-color image. (b) PTFE (Stage-I). (c) PTFE (Stage-II). (d) PTFE (Stage-III). (e) SAFF. (f) Detection map.

S^2TNet, where the false-color map is mixed with the heat map using a ratio of 0.8 and 1. It can be seen that, in the different encoding stages of the PTFE module, the intensity of the heatmap in the target region is the highest, capturing reliable fine-grained target information. Moreover, with the constraint of the sparse self-attention mechanism in the decoding stages, the CAM of the proposed S^2TNet method shows a minimal response in the background areas, accurately providing more precise target localization information. Moreover, the CAM demonstrates that the S^2TNet network accurately outlines target boundaries, a capability enhanced by the effective random masking strategy. The strategy works by randomly extending boundaries to produce a variety of target shapes, thereby enabling the network to accurately detect targets with diverse types and shapes.

2) Network Structure: To validate the effectiveness of the BAC and TKSA modules designed within the S^2TNet network, we analyzed the target detection performance using four different network structures. As shown in Table VI, $\text{AUC}_{(P_d,P_f)}$ and $\text{AUC}_{(P_f,\tau)}$ are used to evaluate the detection performance of the models. It can be observed that adding either BAC or TKSA to the baseline model improves the $\text{AUC}_{(P_d,P_f)}$ score and reduces the $\text{AUC}_{(P_f,\tau)}$ score, thereby confirming the effectiveness of the BAC and TKSA modules. Furthermore, incorporating both BAC and TKSA modules simultaneously enables BASTNet to achieve an optimal performance, demonstrating its superiority in target detection and background suppression.

3) Loss Function: To verify the effectiveness of the proposed BAC module, ablation experiments were conducted for both the background guidance and foreground guidance loss functions. The AUC scores for the five datasets using different combinations of loss components are presented in Table VII. L_{BG} represents the background guidance loss, which is the sum of L_{BL} and L_{BR}. The results indicate that the low-rank background and background ratio losses are crucial for the proposed BAC module, as they effectively suppress background interference. The primary reason is that the L_{BG} loss minimizes the background response, thereby enhancing the detection probability P_d. When the L_{FG} loss is applied, this leads to a significant improvement in $\text{AUC}_{(P_d,P_f)}$, which helps the detector to further enhance its performance. This improvement occurs because L_{FC} and L_{FD} are designed to focus on target regions at different scales, facilitating more accurate target identification. In summary, by jointly constraining the foreground guidance and background guidance losses, the BAC module in BASTNet helps the model achieve an optimal performance.

4) Data Augmentation Strategy: As shown in Fig. 15(a) and (b), nine parameter combinations were designed to assess the influence of the target number range $[N_{\min}, N_{\max}]$ and target area range $[A_{\min}, A_{\max}]$ in the random masking strategy. BASTNet exhibits a downtrend in detection performance as $N_{\max}$ increases, while maintaining a smaller $N_{\min}$ ensures consistent detection capabilities on all five datasets. That is, restricting the number of targets to the range [3], [5] demonstrates strong generalization capability. Appropriately increasing $A_{\max}$ can improve the BASTNet's adaptability to the

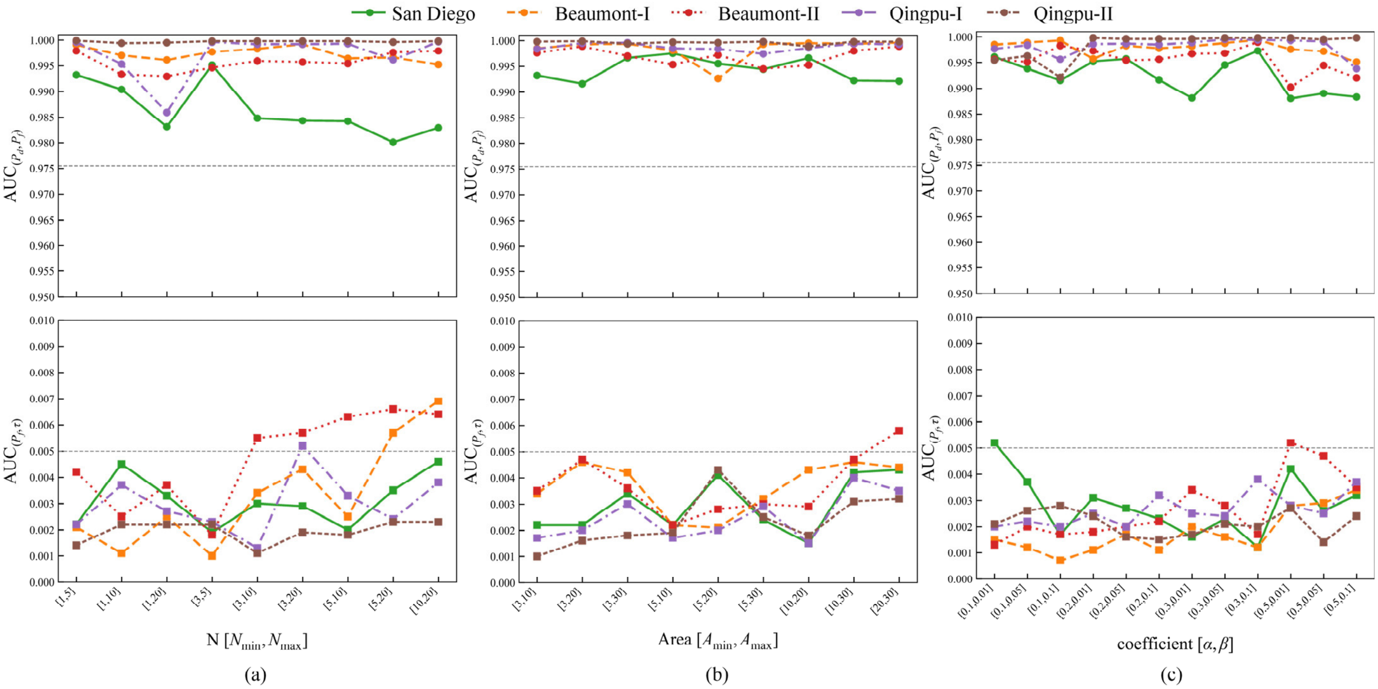


Fig. 15. Impact of data augmentation strategy on target detection performance. (a) Target number range [$N_{\min}, N_{\max}$]. (b) Target area range [$A_{\min}, A_{\max}$]. (c) Parameter combinations [α, β].

TABLE VIII

AUC$_{(P_d,P_f)}$ AND AUC$_{(P_f,\tau)}$ SCORES FOR DIFFERENT TRAINING DATA RATIOS ON THE FIVE DATASETS

Training data ratio	San Diego	Beaumont-I	Beaumont-II	Qingpu-I	Qingpu-II
20%	0.9967/0.0048	0.9990/0.0048	0.9932/0.0067	0.9996/0.0096	**0.9999**/0.0079
40%	0.9947/0.0040	**0.9994**/0.0042	0.9968/0.0047	0.9996/0.0045	0.9998/0.0043
60%	0.9971/0.0033	0.9991/0.0030	0.9981/0.0034	0.9997/0.0021	0.9998/0.0016
80%	0.9970/**0.0010**	0.9991/0.0027	0.9983/0.0020	0.9996/0.0018	0.9998/0.0014
100%	**0.9973**/0.0012	**0.9994/0.0012**	**0.9990/0.0017**	**0.9998/0.0015**	**0.9999/0.0009**

variation in target size, reflecting the diversity encountered in real-world scenarios. Setting a high $A_{\min}$ hinders BASTNet's ability to learn the characteristics of small targets, thus $A_{\min}$ should remain a minimal value to preserve sensitivity to small-scale targets. To summarize, a target area range of [3] and [10] provides an optimal balance between capturing scale diversity and maintaining sensitivity to small-scale targets.

Moreover, we investigated the performance of the model with various settings of the background abundance coefficient α and noise adjustment coefficient β in (3) by establishing twelve sets of parameter combinations [α, β]. As shown in Fig. 15(c), the best AUC is achieved when $\alpha = 0.3$, where the generated background disturbance effectively maintains data authenticity without overshadowing the target signal. The increasing β indicates that an appropriate noise is essential for enhancing the generalization ability of the BASTNet. When α is 0.3 and β is 0.05, the proposed algorithm maintains high detection accuracy and consistent performance.

5) Training Data Ratio: To illustrate the impact of different training samples on the model accuracy, we used 20%, 40%, 60%, 80%, and 100% of the training samples for the training. Table VIII lists the test accuracy of BASTNet on the five datasets under different proportions of training data. It can be observed that, as the training data ratio increases, both AUC$_{(P_d,P_f)}$ and AUC$_{(P_f,\tau)}$ gradually reach optimal values, indicating that a greater amount of training data enhances the detection performance of BASTNet. The most notable trend is the inverse relationship between the sample ratio and AUC$_{(P_f,\tau)}$, where the most significant changes can be observed. In summary, the more background training samples available, the better the network can learn the background, leading to a superior target detection state.

6) Parameter k of TKSA: As shown in Fig. 16, the detection performance of C-TKSA and S-TKSA is evaluated across different k-value ranges [Δ_1, Δ_2]. For C-TKSA, the AUC value decreases when the lower bound for sparsity Δ_1 is set to 1/4, since the BASTNet's ability to distinguish targets is weakened because of the insufficient channel numbers. In contrast, the detection capability of BASTNet is gradually improved as the upper bound Δ_2 increases. Consequently, C-TKSA optimizes the k value in the range of 50%–80% to prevent the loss of high-frequency features that are essential for effective target detection. On the other hand, S-TKSA retains only a limited number of significant feature points when k is small (1/200–1/50), leading to the loss of crucial local target details. In comparison, S-TKSA achieves an optimal balance between preserving target features and reducing background interference by dynamically adjusting the k value to 1%–10%.

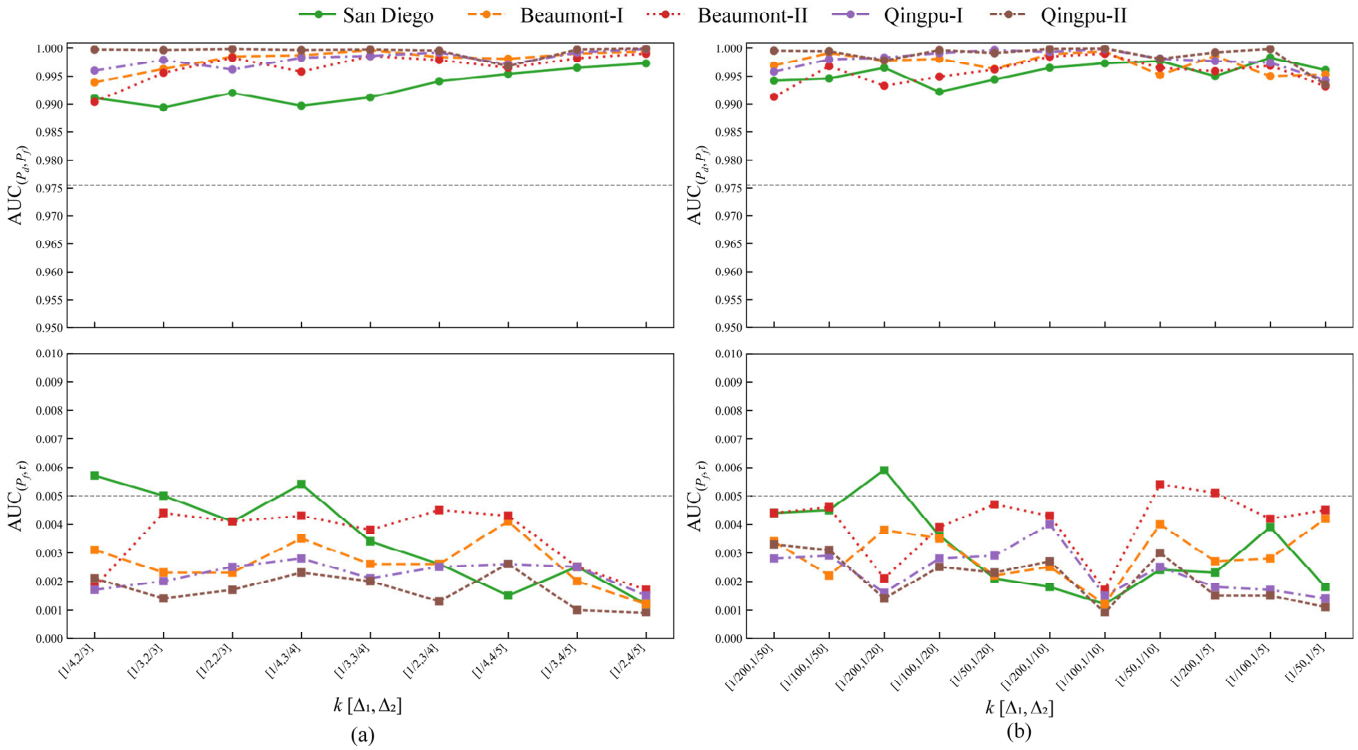


Fig. 16. Impact of TKSA on target detection performance. (a) Different k value ranges of C-TKSA. (b) Different k value ranges of S-TKSA.

V. Conclusion

In this study, we conducted research on training sample construction, model design, and loss formulation, and developed a BASTNet for HTD. First, considering the limitations of training data and the complexity of the background, random masking and spectral generation strategies are employed to generate sufficient and diverse training samples. This image-level data augmentation approach motivates the network to learn the spatial and contextual features of the target. An S^2TNet incorporating the PTFE, SAFF, and TPD modules is applied to capture detailed target information at different scales. In addition, with the top-k operator, the model focuses more on the target regions, to improve the target detection accuracy. Furthermore, we introduce background-aware learning and a novel foreground and background guidance loss function to enhance the separability between target and background. The experimental results confirmed that BASTNet can achieve a superior detection performance and enables rapid inference, compared to the current target detection methods, on five large-scale hyperspectral datasets.

The following will be our future directions for improvement. First, it will be worth exploring the design of an unsupervised automatic background patch segmentation mechanism that also supports the embedding of multiple types of targets, to simplify the data preprocessing process. In addition, we aim to explore a transferable and interpretable deep learning model to extend the practical applicability of the model.

Acknowledgment

The authors would like to extend their sincere gratitude to the Shanghai Institute of Surveying and Mapping for their provision of the airborne hyperspectral imagery data.

References

[1] H. Yu, Z. Ling, K. Zheng, L. Gao, J. Li, and J. Chanussot, "Unsupervised hyperspectral and multispectral image fusion with deep spectral–spatial collaborative constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534114, doi: 10.1109/TGRS.2024.3472226.

[2] R. Ji et al., "PatchOut: A novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 138, Apr. 2025, Art. no. 104457, doi: 10.1016/j.jag.2025.104457.

[3] B. Xi et al., "MCTGCL: Mixed CNN–transformer for Mars hyperspectral image classification with graph contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5503214, doi: 10.1109/TGRS.2025.3529996.

[4] D. Zhu, B. Du, and L. Zhang, "Learning single spectral abundance for hyperspectral subpixel target detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 10134–10144, Jul. 2024, doi: 10.1109/TNNLS.2023.3239061.

[5] W. Dong et al., "Deep spatial–spectral joint-sparse prior encoding network for hyperspectral target detection," *IEEE Trans. Cybern.*, early access, Dec. 5, 2024, doi: 10.1109/TCYB.2024.3403729.

[6] B. Tu, X. Yang, B. He, Y. Chen, J. Li, and A. Plaza, "Anomaly detection in hyperspectral images using adaptive graph frequency location," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 7, pp. 12565–12579, Jul. 2024, doi: 10.1109/TNNLS.2024.3449573.

[7] S. Sun, J. Liu, Z. Zhang, and W. Li, "Hyperspectral anomaly detection based on adaptive low-rank transformed tensor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9787–9799, Jul. 2024, doi: 10.1109/TNNLS.2023.3236641.

[8] B. Tu, X. Yang, W. He, J. Li, and A. Plaza, "Hyperspectral anomaly detection using reconstruction fusion of quaternion frequency domain analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8358–8372, Jun. 2024, doi: 10.1109/TNNLS.2022.3227167.

[9] K. Tan, L. Chen, H. Wang, Z. Liu, J. Ding, and X. Wang, "Estimation of the distribution patterns of heavy metal in soil from airborne hyperspectral imagery based on spectral absorption characteristics," *J. Environ. Manage.*, vol. 347, Dec. 2023, Art. no. 119196, doi: 10.1016/j.jenvman.2023.119196.

[10] N. M. Nasrabadi, "Hyperspectral target detection: An overview of current and future challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 34–44, Jan. 2014, doi: 10.1109/MSP.2013.2278992.

[11] C.-I. Chang, "Constrained energy minimization (CEM) for hyperspectral target detection: Theory and generalizations," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5522921, doi: 10.1109/TGRS.2024.3424281.

[12] F. A. Kruse et al., "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 145–163, May 1993, doi: 10.1016/0034-4257(93)90013-n.

[13] W. Farrand, "Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho, through the use of a constrained energy minimization technique," *Remote Sens. Environ.*, vol. 59, no. 1, pp. 64–76, Jan. 1997, doi: 10.1016/s0034-4257(96)00080-6.

[14] C.-I. Chang, "Orthogonal subspace projection (OSP) revisited: A comprehensive study and analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 502–518, Mar. 2005, doi: 10.1109/TGRS.2004.839543.

[15] Z. Zou and Z. Shi, "Hierarchical suppression method for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 330–342, Jan. 2016, doi: 10.1109/TGRS.2015.2456957.

[16] R. Zhao, Z. Shi, Z. Zou, and Z. Zhang, "Ensemble-based cascaded constrained energy minimization for hyperspectral target detection," *Remote Sens.*, vol. 11, no. 11, p. 1310, Jun. 2019, doi: 10.3390/rs11111310.

[17] X. Yang, M. Zhao, T. Gao, J. Chen, and J. Zhang, "Multiscale-superpixel-based SparseCEM for hyperspectral target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3079445.

[18] X. Zhao, Z. Hou, X. Wu, W. Li, P. Ma, and R. Tao, "Hyperspectral target detection based on transform domain adaptive constrained energy minimization," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102461, doi: 10.1016/j.jag.2021.102461.

[19] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 629–640, Jun. 2011, doi: 10.1109/JSTSP.2011.2113170.

[20] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, Dec. 2015, doi: 10.1016/j.patcog.2015.05.024.

[21] D. Zhu, B. Du, M. Hu, Y. Dong, and L. Zhang, "Collaborative-guided spectral abundance learning with bilinear mixing model for hyperspectral subpixel target detection," *Neural Netw.*, vol. 163, pp. 205–218, Jun. 2023, doi: 10.1016/j.neunet.2023.02.002.

[22] T. Cheng and B. Wang, "Decomposition model with background dictionary learning for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1872–1884, 2021, doi: 10.1109/JSTARS.2021.3049843.

[23] C. Li, D. Zhu, C. Wu, B. Du, and L. Zhang, "Global overcomplete dictionary-based sparse and nonnegative collaborative representation for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5513214, doi: 10.1109/TGRS.2024.3381719.

[24] X. Nie, Z. Xue, C. Lin, L. Zhang, and H. Su, "Structure-prior-constrained low-rank and sparse representation with discriminative incremental dictionary for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5506319, doi: 10.1109/TGRS.2024.3353370.

[25] S. Feng, R. Feng, D. Wu, C. Zhao, W. Li, and R. Tao, "A coarse-to-fine hyperspectral target detection method based on low-rank tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5530413, doi: 10.1109/TGRS.2023.3329800.

[26] X. Zhao, K. Liu, K. Gao, and W. Li, "Hyperspectral time-series target detection based on spectral perception and spatial–temporal tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5520812, doi: 10.1109/TGRS.2023.3307071.

[27] Y. Wang, X. Chen, E. Zhao, C. Zhao, M. Song, and C. Yu, "An unsupervised momentum contrastive learning based transformer network for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9053–9068, 2024, doi: 10.1109/JSTARS.2024.3387985.

[28] D. Shen et al., "HTD-Mamba: Efficient hyperspectral target detection with pyramid state space model," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5507315, doi: 10.1109/TGRS.2025.3547019.

[29] G. Zhang, S. Zhao, W. Li, Q. Du, Q. Ran, and R. Tao, "HTD-Net: A deep convolutional neural network for target detection in hyperspectral imagery," *Remote Sens.*, vol. 12, no. 9, p. 1489, May 2020, doi: 10.3390/rs12091489.

[30] Y. Wang, X. Chen, E. Zhao, and M. Song, "Self-supervised spectral-level contrastive learning for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510515, doi: 10.1109/TGRS.2023.3270324.

[31] L. Sun, Z. Ma, and Y. Zhang, "ABLAL: Adaptive background latent space adversarial learning algorithm for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 411–427, 2024, doi: 10.1109/JSTARS.2023.3329771.

[32] Q. Tian, C. He, Y. Xu, Z. Wu, and Z. Wei, "Hyperspectral target detection: Learning faithful background representations via orthogonal subspace-guided variational autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5516714, doi: 10.1109/TGRS.2024.3393931.

[33] J. Jiao, Z. Gong, and P. Zhong, "Triplet spectralwise transformer network for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5519817, doi: 10.1109/TGRS.2023.3306084.

[34] S. Feng et al., "Transformer-based cross-domain few-shot learning for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5501716, doi: 10.1109/TGRS.2024.3521035.

[35] H. Qin, W. Xie, Y. Li, K. Jiang, J. Lei, and Q. Du, "Weakly supervised adversarial learning via latent space for hyperspectral target detection," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109125.

[36] W. Xie, J. Zhang, J. Lei, Y. Li, and X. Jia, "Self-spectral learning with GAN based spectral–spatial target detection for hyperspectral image," *Neural Netw.*, vol. 142, pp. 375–387, Oct. 2021, doi: 10.1016/j.neunet.2021.05.029.

[37] J. Lei, M. Li, W. Xie, Y. Li, and X. Jia, "Spectral mapping with adversarial learning for unsupervised hyperspectral change detection," *Neurocomputing*, vol. 465, pp. 71–83, Nov. 2021, doi: 10.1016/j.neucom.2021.08.130.

[38] X. Chen, Y. Zhang, Y. Dong, and B. Du, "Generative self-supervised learning with spectral–spatial masking for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5522713, doi: 10.1109/TGRS.2024.3423781.

[39] F. Luo, S. Shi, K. Qin, T. Guo, C. Fu, and Z. Lin, "SelfMTL: Self-supervised meta-transfer learning via contrastive representation for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5508613, doi: 10.1109/TGRS.2025.3550283.

[40] W. Xie, X. Zhang, Y. Li, K. Wang, and Q. Du, "Background learning based on target suppression constraint for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5887–5897, 2020, doi: 10.1109/JSTARS.2020.3024903.

[41] D. Shen, X. Ma, W. Kong, J. Liu, J. Wang, and H. Wang, "Hyperspectral target detection based on interpretable representation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5519416, doi: 10.1109/TGRS.2023.3302950.

[42] H. Qin, S. Wang, Y. Li, W. Xie, K. Jiang, and K. Cao, "Hyperspectral target detection based on generative self-supervised learning with wavelet transform," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, doi: 10.1109/TGRS.2025.3549771.

[43] H. Qin, S. Wang, Y. Li, W. Xie, K. Jiang, and K. Cao, "A signature-constrained two-stage framework for hyperspectral target detection based on generative self-supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5514917, doi: 10.1109/TGRS.2025.3564039.

[44] X. Chen, Y. Zhang, Y. Dong, and B. Du, "Spatial–spectral contrastive self-supervised learning with dual path networks for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5515612, doi: 10.1109/TGRS.2024.3390946.

[45] Y. Gao, Y. Feng, and X. Yu, "Hyperspectral target detection with an auxiliary generative adversarial network," *Remote Sens.*, vol. 13, no. 21, p. 4454, Nov. 2021, doi: 10.3390/rs13214454.

[46] Z. Li et al., "Hyperspectral target detection using diffusion model and convolutional gated linear unit," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5512313, doi: 10.1109/TGRS.2025.3565361.

[47] P. Zhong, Z. Gong, and J. Shan, "Multiple instance learning for multiple diverse hyperspectral target characterizations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 246–258, Jan. 2020, doi: 10.1109/TNNLS.2019.2900465.

[48] H. Gao, Y. Zhang, Z. Chen, F. Xu, D. Hong, and B. Zhang, "Hyperspectral target detection via spectral aggregation and separation network with target band random mask," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515516, doi: 10.1109/TGRS.2023.3288739.

[49] C. Li, R. Wang, Z. Chen, H. Gao, and S. Xu, "Transformer-inspired stacked-GAN for hyperspectral target detection," *Int. J. Remote Sens.*, vol. 45, no. 15, pp. 4961–4982, Aug. 2024, doi: 10.1080/01431161.2024.2370500.

[50] H. Qin, W. Xie, Y. Li, and Q. Du, "HTD-TS^3: Weakly supervised hyperspectral target detection based on transformer via spectral–spatial similarity," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 11, pp. 16816–16830, Nov. 2023, doi: 10.1109/TNNLS.2023.3298145.

[51] Y. Li, H. Qin, and W. Xie, "HTDFormer: Hyperspectral target detection based on transformer with distributed learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524715, doi: 10.1109/TGRS.2023.3317033.

[52] Q. Yang, X. Wang, L. Chen, Y. Zhou, and S. Qiao, "CS-TTD: Triplet transformer for compressive hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5533115, doi: 10.1109/TGRS.2024.3436084.

[53] Y. Shi, H. Cui, Y. Yin, H. Song, Y. Li, and P. Gamba, "Transfer learning with nonlinear spectral synthesis for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5532517, doi: 10.1109/TGRS.2023.3336688.

[54] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data augmentation for GAN training," *IEEE Trans. Image Process.*, vol. 30, pp. 1882–1897, 2021, doi: 10.1109/TIP.2021.3049346.

[55] Z. Zhuang, J. Lan, and Y. Zeng, "Hyperspectral target detection based on masked autoencoder data augmentation," *Remote Sens.*, vol. 17, no. 6, p. 1097, Mar. 2025, doi: 10.3390/rs17061097.

[56] Z. Li, Y. Wang, C. Xiao, Q. Ling, Z. Lin, and W. An, "You only train once: Learning a general anomaly enhancement network with random masks for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506718, doi: 10.1109/TGRS.2023.3258067.

[57] J. Li, X. Wang, S. Wang, H. Zhao, and Y. Zhong, "One-step detection paradigm for hyperspectral anomaly detection via spectral deviation relationship learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5517515, doi: 10.1109/TGRS.2024.3392189.

[58] Y. Li, K. Jiang, W. Xie, J. Lei, X. Zhang, and Q. Du, "A model-driven deep mixture network for robust hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522916, doi: 10.1109/TGRS.2023.3309960.

[59] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412, doi: 10.1109/TGRS.2023.3279834.

[60] X. Zhang et al., "Self-supervised learning with deep clustering for target detection in hyperspectral images with insufficient spectral variation prior," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103405, doi: 10.1016/j.jag.2023.103405.

[61] W. Rao, L. Gao, Y. Qu, X. Sun, B. Zhang, and J. Chanussot, "Siamese transformer network for hyperspectral image target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526419, doi: 10.1109/TGRS.2022.3163173.

[62] H. A. Amirkolaee, M. Shi, and M. Mulligan, "TreeFormer: A semi-supervised transformer-based framework for tree counting from a single high-resolution image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4406215, doi: 10.1109/TGRS.2023.3295802.

[63] Q. Zhang and Y.-B. Yang, "ResT: An efficient transformer for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15475–15485.

[64] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5896–5905.

[65] R. Dian, Y. Liu, and S. Li, "Spectral super-resolution via deep low-rank tensor representation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 3, pp. 5140–5150, Mar. 2024, doi: 10.1109/TNNLS.2024.3359852.

[66] W. Zhai, P. Wu, K. Zhu, Y. Cao, F. Wu, and Z.-J. Zha, "Background activation suppression for weakly supervised object localization and semantic segmentation," *Int. J. Comput. Vis.*, vol. 132, no. 3, pp. 750–775, Mar. 2024, doi: 10.1007/s11263-023-01919-2.

[67] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[68] Y. Zhang, K. Wu, B. Du, and X. Hu, "Multitask learning-based reliability analysis for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2135–2147, Jul. 2019, doi: 10.1109/JSTARS.2019.2894802.

[69] C.-I. Chang, "An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5131–5153, Jun. 2021, doi: 10.1109/TGRS.2020.3021671.

Zhiwei Wang received the B.S. degree in surveying and mapping engineering and the M.S. degree in photogrammetric and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in cartography and geographic information systems with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China.

His research interests include hyperspectral image processing, anomaly detection, and object detection.

Kun Tan (Senior Member, IEEE) received the B.S. degree in information and computer science from Hunan Normal University, Changsha, China, in 2004, and the Ph.D. degree in photogrammetric and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2010.

From 2008 to 2009, he was a Joint Ph.D. Candidate of remote sensing with Columbia University, New York, NY, USA. From 2010 to 2018, he was with the Department of Surveying, Mapping and Geoinformation, China University of Mining and Technology, Xuzhou. He is currently a Professor with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China. His research interests include hyperspectral image classification and detection, spectral unmixing, quantitative inversion of land surface parameters, and urban remote sensing.

Xue Wang received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetric and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2014 and 2019, respectively.

He is currently an Associate Professor with the Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China. His research interests include hyperspectral imagery processing, deep learning, and ecological monitoring.

Xiaojun Xu received the B.S. degree in biological science from Zhejiang Normal University, Jinhua, China, in 2003, and the M.S. degree in ecology from East China Normal University, Shanghai, China, in 2006.

He is currently a Senior Engineer with Shanghai Environmental Monitoring Center. His research interests include the applications of remote sensing and unmanned aerial vehicle (UAV) technology in environmental monitoring.