HTD-VIT: SPECTRAL-SPATIAL JOINT HYPERSPECTRAL TARGET DETECTION WITH VISION TRANSFORMER

Haonan Qin[1]*, Weiying Xie*[1]*, Yunsong Li*[1]*, and Qian Du*[2]

[1]State Key Lab. of Integrated Services Networks, Xidian University, Xi'an, China
[2]The Department of Electrical and Computer Engineering, Mississippi State University, USA.

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884695

ABSTRACT

In hyperspectral images (HSIs), spatial context provides complementary information to abundant spectral features. In this paper, a united spectral-spatial framework named HTD-ViT based on vision transformer (ViT) is proposed for HTD tasks. The HTD-ViT leverages the ViT to learn discriminative spectral-spatial features of each pixel and its neighboring pixels. Meanwhile, the spectral-spatial sequence construction operation uses spectrums in the cross region centered on the selected pixel to produce the corresponding spectral-spatial sequence for ViT processing. Furthermore, the spectral-spatial sample selection procedure based on coarse detection addresses the issue of lacking well-labeled training instances in the HTD tasks. Finally, the spectral-spatial pixel-level detection combines the discriminative feature from the spectral and the spatial domains to suppress the background. In contrast to traditional spatial-spectral feature extraction methods that stack the original spectral feature with spatial neighborhood information directly, joint spectral-spatial inference in HTD-ViT can effectively discover the underlying contextual and structure information in HSIs. Experiments on real HSIs verify the effectiveness of HTD-ViT, which takes full advantage of both the variable spectral and spatial features.

Index Terms— Hyperspectral target detection, vision transformer, spectral-spatial framework

1. INTRODUCTION

Hyperspectral target detection (HTD) automatically extracts targets of interest from hyperspectral image (HSI) [1]. Recently, many methods have been proposed such as CEM [2], ACE [3], CSCR [4], hCEM [5] and E-CEM [6]. These methods have shown different advantages and improvements [1].

However, there are still some challenges for the HTD tasks: (1) Inadequate use of the spectral and spatial information restrains the performance of the detector. (2) Limited availability of labeled samples makes it hard for the detector to learn the distribution of real HSIs.(3) Hundreds of contiguous spectral bands of HSIs prevent machine learning methods from being directly transferred to hyperspectral tasks.

In reality, the spatial contexts of HSIs can provide complementary information to their abundant spectral features for more precise recognition. So it is significant to make full use of the correlation between the spectral and spatial domains.

We are inspired by the ideas of the vision transformer (ViT) [7] which successfully utilizes the transformer [8] for vision-based tasks. The ViT structure realizes flexible learned allocation of global attention in an image. On this basis, we propose a novel architecture for spectral-spatial joint hyperspectral target detection with vision transformer (HTD-ViT). HTD-ViT extracts the spectral signatures and spatial features jointly in a united spectral-spatial framework. It consists of four spectral-spatial procedures: sample selection, sequence construction, transformer training and inference fusion.

Compared with the above methods, the proposed joint spectral-spatial framework shows that the learned spectral-spatial feature is more discriminative for HTD. And it avoids designing the artificial parameters that are sensitive to the local changes of the input data, especially when the training samples are limited for HTD. Contributions are as follows.

(1) A united spectral-spatial framework is proposed to automatically learn the joint spectral-spatial features from real HSIs. The joint spectral-spatial features are more discriminative than handcrafted features in traditional methods.

(2) The ViT mechanism is firstly embedded into the HTD tasks to learn global spectral-spatial features of spectrums.

(3) An universal form designed to produce spectral-spatial sequence for HTD can use spectrums of each pixel and its neighboring pixels to produce sequences as the input of ViT.

(4) The spectral-spatial sample selection and inference fusion operation make the spectral and the spatial features extracted and fused for higher-level spectral-spatial inference.

2. PROPOSED APPROACH

2.1. Spectral-Spatial Sample Selection

To cope with the lacking of extremely limited labeled samples for HTD, the sample selection procedure based on the coarse

This work was supported by the National Natural Science Foundation of China under Grant 62121001, Grant 62071360, and Grant 61801359. Additionally, it was also supported by the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University.

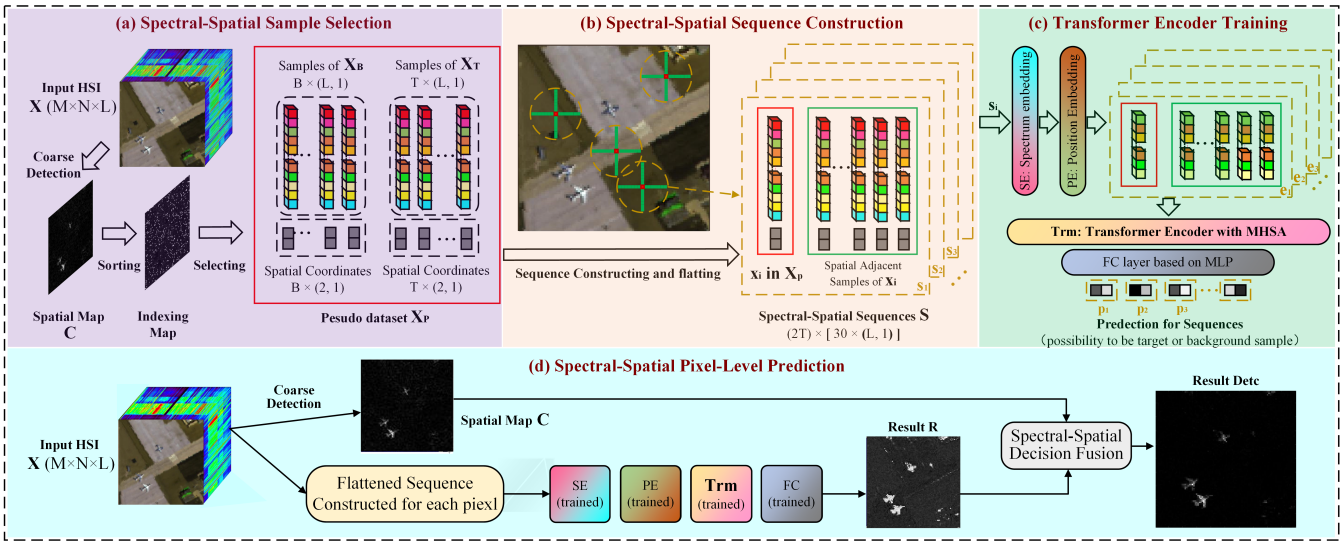

Fig. 1. The HTD-ViT framework consists of four procedures for spectral-spatial feature extraction, information fusion and prediction inference: (1) sample selection, (2) sequence construction, (3) transformer encoder training, (4) pixel-level prediction.

detection is proposed. Let $\mathbf{X} \in \mathbf{R}^{M\times N\times L}$ denote the input HSI with $M \times N$ pixels, L spectral bands and the target signature $\mathbf{d}$. As shown in Fig.1 (a), we sort the pixels of $\mathbf{X}$ based on the value of each pixel of the coarse detection result $\mathbf{C}$ via the widely used CEM detector [2] as a spatial reference. $\mathbf{C}$ can be regarded as an initial spatial map, and each pixel represents the likelihood of a certain spectrum being a target sample. Then, we select the top 30% samples with relatively small values to form pseudo-background dataset $\mathbf{X}_B \in \mathbf{R}^{B\times L}$ where $B = M \times N \times 30\%$. And we select the last 1.5% samples to form the pseudo-target dataset $\mathbf{X}_T \in \mathbf{R}^{T\times L}$ where $T = M \times N \times 1.5\%$. Finally, the spectral-spatial pseudo dataset $\mathbf{X}_P$ composed of spectrums in $\mathbf{X}_B$ and $\mathbf{X}_T$ can be obtained. During training, the T samples in $\mathbf{X}_T$ are all for training in an epoch. And training samples from $\mathbf{X}_B$ are composed of randomly sampled T samples from $\mathbf{X}_B$. Considering pseudo labels may be inaccurate and incomplete, HTD-ViT can be regarded as a weakly supervised learning method.

2.2. Spectral-Spatial Sequence Construction

To ensure that the training and test samples meet the input requirement of the transformer encoder, HTD-ViT employs the spectral-spatial sequence construction procedure. In ViT [7], an image will be split into patches and provide the sequence of these patches as an input to a transformer encoder. These image patches are treated the same way as words in an NLP application [8]. Considering 3-D HSIs can be regarded as 3-D tensors, the sequence construction rule based on spatial position information has been developed in this work.

Specifically, given a pixel $\mathbf{x}_i$ in $\mathbf{X}_P$ (we want to predict its label), a cross region containing the pixel is selected and flattened into a pixel sequence $\mathbf{s}_i$, which makes the model more efficient than square region strategy. The cross region is composed of a horizontal line and a vertical line, in which both the horizontal line and vertical line contain 15 pixels, respectively. As shown in Fig. 1 (b), a cross region contains 30 pixels in total including the centered pixel $\mathbf{x}$. Therefore, the sequence samples set $\mathbf{S}$ for training HTD-ViT can be obtained based on $\mathbf{X}_P$ and $\mathbf{X}$. Then, each pixel $\mathbf{x}_i$ in $\mathbf{X}_P$ will attend to every pixel in corresponding pixel sequence $\mathbf{s}_i$ through the multi-head self-attention (MHSA) [8] mechanism of transformer.

2.3. Transformer Encoder Training

To take full advantage of the variable spectral and spatial features, ViT with a global receptive field [7] via self-attention layer is firstly applied into HTD. HTD-ViT leverages ViT to capture discriminative spectral-spatial features of each pixel and its neighboring pixels regardless of their spatial distance.

For each pixel $\mathbf{x}_i$ in an input spectral-spatial sequence $\mathbf{s}_i$, the self-attention layer will compute an attention distribution over the cross region, and the attention distribution informs $\mathbf{x}_i$ where it should attend to. Firstly, the pixel sequence $\mathbf{s}_i$ is transformed using the spectrum embedding (SE) operation, which transforms pixel vectors to a new vector space with a predefined dimension D. Then, the position embedding (PE) operation is added to the spectrum embeddings to retain positional information. SE and PE can be formulated as:

$$\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2; \cdots; \mathbf{e}_n] = [\mathbf{s}_1\mathbf{E}_\mathrm{S}; \mathbf{s}_2\mathbf{E}_\mathrm{S}; \cdots; \mathbf{s}_n\mathbf{E}_\mathrm{S}] + \mathbf{E}_\mathrm{P} \quad (1)$$

where $\mathbf{E}_\mathrm{S}$ represents the learned weights of the linear transformation, and $\mathbf{E}_\mathrm{P}$ is the learned positional matrix.

Therefore, for the spectral-spatial sequences $\mathbf{S}$, the SE and PE operation allow transformer encoder with the desired architecture can be used in all input dimensions. Besides, the

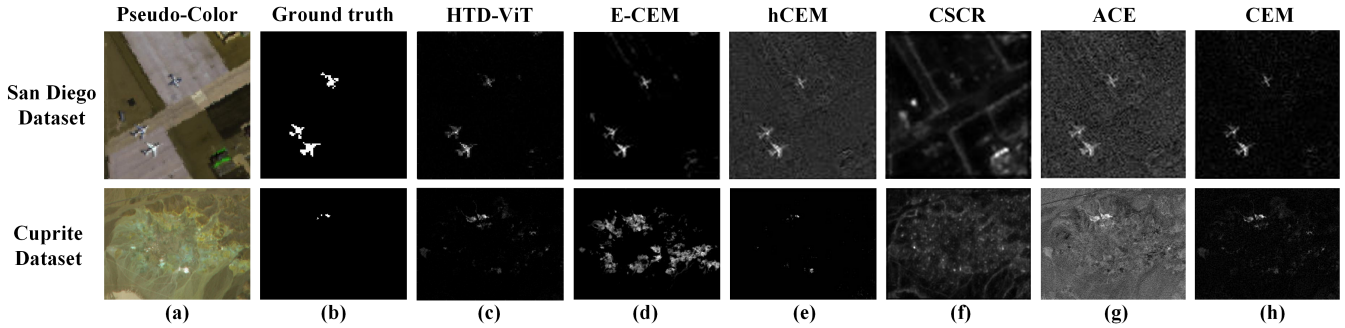

Fig. 2. Datasets for test and detection results on the San Diego dataset and the Cuprite dataset of the six compared methods.

two operations can fasten the training speed by reducing the input dimension of HSIs. And the resulting sequences of embedding vectors $\mathbf{E}$ serve as input to the transformer encoder.

Then, the transformer encoder in HTD-ViT employs MHSA to learn multiple relations among the pixels in a pixel sequence, with each head learning different relations separately and in parallel. Specifically, MHSA can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weights assigned to each value will be computed by a compatibility function of the query with the corresponding key [8].

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query matrix, key matrix, and value matrix, respectively, and d is the dimension of the input data. And MHSA with m heads can be formulated as:

$$\text{MHSA}(\mathbf{E})=\text{Concat}\left(\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_m\right)\mathbf{W}^{O} \tag{3}$$

where $\mathbf{h}_i=\text{Attention}\left(\mathbf{E}\mathbf{W}_i^{Q},\mathbf{E}\mathbf{W}_i^{K},\mathbf{E}\mathbf{W}_i^{V}\right)$, $\mathbf{W}$ represent the corresponding learned matrices, and $\mathbf{e}_i\in\mathbf{E}$.

The layernorm (LN) [7] operation is applied after MHSA:

$$\mathbf{e}_i^{Trm}=\text{LN}\left(\text{MHSA}(\mathbf{e}_i)\right) \tag{4}$$

Finally, the features learned via transformer encoder are fed into a single-layer fully connected (FC) neural network with two linear transformations with a ReLU activation.

$$\mathbf{f}_i=\text{LN}\left(\max\left(W_1\mathbf{e}_i^{Trm}+b_1,0\right)W_2+b_2\right) \tag{5}$$

Then, the pixel-level prediction for the possibility of the training sample $\mathbf{x}_i$ to be a target (belong to $\mathbf{X}_T$) or a background sample (belong to $\mathbf{X}_B$) can be obtained by softmax operation:

$$\mathbf{p}_i=\text{softmax}\left(\mathbf{f}_i\right) \tag{6}$$

The loss value in an epoch in training can be computed as:

$$Loss=\frac{1}{2T}\sum_i-\left[y_i\cdot\log\left(p_i^1\right)+\left(1-y_i\right)\cdot\log\left(p_i^2\right)\right] \tag{7}$$

where p_i^1 and p_i^2 represent the prediction possibility of sample x_i to be target and background, respectively.

And y_i represents the pseudo label of sample x_i as:

$$y_i=\begin{cases}1, & \mathbf{x}_i\in\mathbf{X}_T\\ 0, & \mathbf{x}_i\in\mathbf{X}_B\end{cases} \tag{8}$$

2.4. Spectral-Spatial Pixel-Level Prediction

After training, as shown in Fig.1 (d), each pixel of the HSI for test can be processed by sequence construction to produce test sequences. The sequences will be processed by SE, PE and trained transformer encoder to realize pixel-level prediction and produce prediction $\mathbf{R}$. Considering the extracted spectral information and spatial information are complementary to each other, spectral-spatial infernence fusion is proposed.

HTD-ViI imposes the joint spectral-spatial inference result $\mathbf{R}$ on the spatial map $\mathbf{C}$ to combine the discriminative feature from the spectral and spatial domains to suppress the background. The final detection result $\mathbf{Detc}\in\mathbf{R}^{M\times N}$ is produced by exponential constrained nonlinear transform as:

$$\mathbf{Detc}=\left(\mathbf{1}-e^{-0.05\cdot\beta\cdot\mathbf{C}}\right)\odot\mathbf{R} \tag{9}$$

where β is an adjustable parameter for different HSIs.

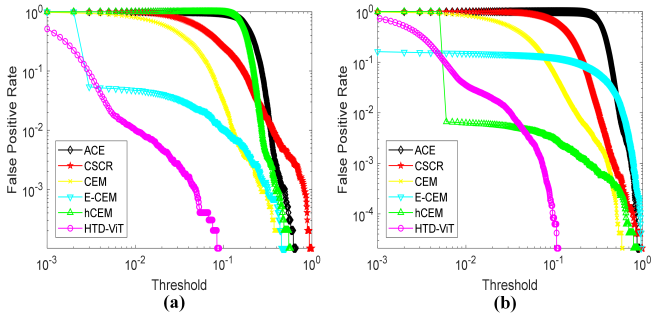

Fig. 3. ROC curves on (a) San Diego, (b) Cuprite. The closer to (0, 0) of a curve, the better corresponding method performs.

3. EXPERIMENT AND ANALYSIS

3.1. Datasets and Experimental Setting

All the experiments have been carried out in the TensorFlow framework on an Intel Core-i7 PC with 16GB of RAM. Two real HSIs captured by the AVIRIS sensor are utilized: the San Diego dataset having 100×100 pixels with 189 bands and the Cuprite dataset having 250×191 pixels with 188 bands. There are 3 airplanes occupying 134 pixels and 14 kinds of mineral occupying 41 pixels can be regarded as targets in the San Diego dataset and the Cuprite dataset, respectively. For compared methods, the CEM [2], ACE [3], CSCR [4], hCEM [5] and E-CEM [6] are adopted as a comparison. For setting, we train HTD-ViT for 200 iterations. The learning rate is set as 3×10^{-3} during training. For all datasets, the predefined dimension D and the number of heads in MHSA m are set as 50 and 1, respectively. The adjustable parameter β is set as 5 for both the San Diego dataset and the Cuprite dataset.

3.2. Detection Performance Analysis

In this work, the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC curve (AUC) are adopted for evaluation [1]. We use AUC1 and AUC2 to refer to the AUC value of (FPR, TPR) and (FPR, threshold), respectively. Some detection instances, the ROC curves and corresponding AUC values are shown in Fig. 2, Fig. 3 and Table 1. From Fig. 2, HTD-ViT can detect different targets in different scenes. Besides, it not only effectively removes the redundant and irrelevant background information, but also retains the target structure information, which outperforms other compared methods. From Table 1, the AUC2 value can evaluate the effect of background interference suppression. HTD-ViT achieves the smallest AUC2 values on the two real HSIs, which shows it reduces the false alarm rate effectively to avoid false detection. And the AUC1 values of HTD-ViT are the highest on the San Diego dataset and comparative on the Cuprite dataset, which indicates different targets can be detected well. Therefore, HTD-ViT performs relatively better than others. Moreover, HTD-ViT seems more effective in the San Diego dataset. Considering airplane has more spatial geometric features than mineral, HTD-ViT makes full use of spatial features in HSIs for spectral-spatial joint inference. In the future, HTD-ViT will be tested and optimized for more scenes and targets clssses with various spatial characteristics.

Methods	San Diego Dataset		Cuprite Dataset	
	AUC_1	AUC_2	AUC_1	AUC_2
CEM	0.96687	0.03424	**0.99994**	0.04110
ACE	0.97327	0.19438	0.99982	0.34652
CSCR	0.86643	0.08117	0.89039	0.12936
hCEM	0.98153	0.17481	0.92550	0.00638
E-CEM	0.99158	0.00595	0.98492	0.04280
HTD-ViT	**0.99356**	**0.00157**	0.99990	**0.00330**

Table 1. The AUC scores of six methods on two datasets.

4. CONCLUSION

Aiming at extracting spectral-spatial features for joint inference, we propose HTD-ViT. As a united spectral-spatial framework based on ViT, HTD-ViT can make full use of the discriminative spectral-spatial features of each pixel and its neighboring pixels. Spectral-spatial sample selection and sequence construction address the issue of lacking well-labeled instances for ViT training. Finally, pixel-level detection based on the trained ViT and the coarse detection combines features from the spectral domain and the spatial domain. Experiments on real HSIs verify the effectiveness of HTD-ViT.

5. REFERENCES

[1] Haonan Qin, Weiying Xie, Yunsong Li, Kai Jiang, Jie Lei, and Qian Du, "Ptgan: A proposal-weighted two-stage gan with attention for hyperspectral target detection," in *Proceedings of IGARSS*, 2021, pp. 4428–4431.

[2] Harsanyi and Joseph C, "Detection and classification of subpixel spectral signatures in hyperspectral image sequences," 01 1993.

[3] S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 1–16, 2001.

[4] Li Wei, Du Qian, and Zhang Bing, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognition*, vol. 48, no. 12, 2015.

[5] Zhengxia Zou and Zhenwei Shi, "Hierarchical suppression method for hyperspectral target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 330–342, 2015.

[6] Rui Zhao, Zhenwei Shi, Zhengxia Zou, and Zhou Zhang, "Ensemble-based cascaded constrained energy minimization for hyperspectral target detection," *Remote Sensing*, vol. 11, no. 11, pp. 1310, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of ICLR 2021*, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of NIPS*, 2017, pp. 6000–6010.