IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883439

CONTRASTIVE LEARNING FOR HYPERSPECTRAL TARGET DETECTION

Xi Chen, Yulei Wang, Zongwei Che, Liyu Zhu, Meiping Song and Haoyang Yu

Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 116026, China

ABSTRACT

With the development and progress of deep learning, the use of deep learning technology for hyperspectral target detection has achieved excellent results. However, most deep-learning-based methods do not effectively suppress background. This paper presents a contrastive learning-based hyperspectral target detection (CLHTD) for this purpose. The positive and negative pairs are constructed through data augmentation, and the backbone is used to extract the representative vectors of the augmented samples. Then the representative vectors are mapped to the spectral and the cluster contrast space using their corresponding contrastive head, respectively. In the contrast space, the similarity and dissimilarity of spectra and clusters are learned by maximizing the similarity of positive pairs while minimizing the similarity of negative pairs, to increase the difference between the representative vectors of target and background. Finally, the detection result is obtained through the cosine distance. Experimental results illustrate that the proposed CLHTD algorithm can achieve superior performances for hyperspectral target detection.

Index Terms— Target Detection, contrastive learning, hyperspectral Imagery.

1. INTRODUCTION

Hyperspectral image (HSI) obtains the three-dimensional spectral image of the observation scene through the imaging spectrometer, including both spatial and spectral dimensions, and each pixel in the spatial space contains tens to hundreds of bands of spectral information. Given this advantage, target detection has been a major research area in hyperspectral image to effectively identify and distinguish substances by details of spectral information.

Recently, deep learning has been gradually applied to hyperspectral target detection. In [1], Li et al. adopted the idea of transfer learning, using a reference HSI with labeled information to expand the training samples by pairing between same classes of pixels and between different classes of pixels. The training samples were used to train the deep 1D convolutional neural network (CNN), and the trained model of the deep 1D CNN was used to detect the target. In [2], Zhang et al. designed U-AE structure to generate potential target samples based on the idea of U-net. According to the known target samples, the background samples which are significantly different from the target are found by linear prediction algorithm. After pairing the target pixels with the target pixels and the target pixels with the background pixels, the training samples are expanded to train a 16 layers 1D deep CNN. In [3], it constructed the adversarial automatic encoder (AAE) based on the idea of generating adversarial network. The Constraint energy minimization algorithm was used to filter the HSI to obtain the background samples. The background samples were sent into AAE to learn until convergence. The loss function was added with the target suppression constraint loss to suppress the AAE reconstruction target. The HIS to be detected is reconstructed into a new HIS through the trained AAE, and the background reconstruction of the reconstructed HSI is good, but the difference of target reconstruction is large. In [4], a spectral regularization unsupervised network was designed to introduce spectral regularization into the autoencoder (AE) and the variational autoencoder (VAE) to enable hidden nodes to better represent the spectral information in HSI. The specific nodes that can distinguish the target from the background are selected based on the spectral angle difference between the a priori target and the input pixel spectra, and the discriminative map is obtained by adaptively weighting the feature maps output from the selected specific nodes through a structure tensor-based adaptive weighting method and suppressing the background and local smoothing by morphological open operation with guided filtering to obtain the final detection results.

In this paper, a novel hyperspectral target detection algorithm named as CLHTD is proposed to better distinguish the target from the background. As shown in Fig. 1, the proposed CLHTD algorithm trains a backbone that can distinguish spectral similarity-dissimilarity through contrastive learning [5]. Then, the backbone is used to extract representative vectors of the HSI to be detected and the prior target spectrum. Finally, the similarity between the spectrum of each pixel in the HSI to be detected and the spectrum of prior target is judged by cosine similarity according to the representative vector, and the detection result is obtained.

2. PROPOSED TARGET DETECTION METHOD

2.1. Data augmentation

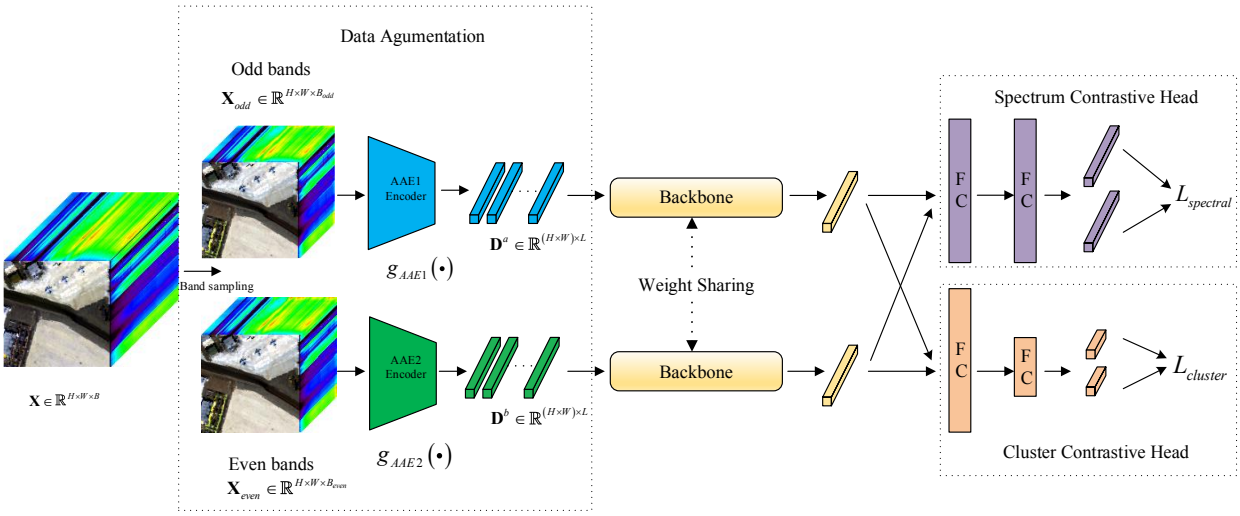


Fig. 1. The framework of the proposed hyperspectral target detection algorithm

Since contrastive learning requires positive and negative sample pairs for training, and the HSI to be detected does not have any other label information except the prior target spectrum, so through data augmentation, two augmented samples of the HSI to be detected are obtained. Two augmented samples of the same pixel spectrum are paired to form a positive pair, and the augmented samples of different pixels are paired to form negative pairs. The data augmentation method is as follows.

Firstly, due to the strong correlation between the adjacent bands of HSI, band sampling is performed on the HSI $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$ to be detected, and two HSI composed of odd and even bands are obtained, denoted as $\mathbf{X}_{odd} \in \mathbb{R}^{H\times W\times B_{odd}}$ and $\mathbf{X}_{even} \in \mathbb{R}^{H\times W\times B_{even}}$, respectively. Then the AAE is trained with $\mathbf{X}_{odd}$ and $\mathbf{X}_{even}$, respectively. The training of AAE includes two parts: the autoencoder network and the adversarial network. During the autoencoder network training phase, the encoder $G_1(\cdot)$ and the decoder $G_2(\cdot)$ make up the autoencoder network. The autoencoder network is optimized by minimizing the reconstruction loss. The reconstruction loss adopts the mean square error loss, which is defined as

$$L_r = \frac{1}{H\times W}\sum_{i=1}^{H\times W}\left\|\mathbf{x}_i - G_2\left(G_1\left(\mathbf{x}_i\right)\right)\right\|_2^2 \tag{1}$$

In the training phase of the adversarial network, the generator $G_1(\cdot)$ (encoder) and discriminator $D(\cdot)$ form the adversarial network. The goal of adversarial training is to make latent code output by generator $G_1(\cdot)$ get closer to prior distribution $p(\mathbf{z})$, while making the discriminator $D(\cdot)$ to better distinguish the feature vector from the latent code output of the generator or the vector sampled from the prior distribution. The prior distribution $p(\mathbf{z})$ is a multivariate Gaussian distribution. The overall optimization goal of the adversarial training can be expressed as

$$\min_{G_1}\max_{D} E_{\mathbf{z}\sim p(\mathbf{z})}\left[\log D(\mathbf{z})\right] + E_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\log\left(1-D\left(G_1(\mathbf{x})\right)\right)\right] \tag{2}$$

When the training is complete, two corresponding encoders $g_{AAE1}(\cdot)$ and $g_{AAE2}(\cdot)$ can be obtained. Regarding $g_{AAE1}(\cdot)$ and $g_{AAE2}(\cdot)$ as transformation functions that play the role of data augmentation. Then using them for data augmentation, the process can be expressed as follows:

$$\mathbf{D}^a = g_{AAE1}\left(\mathbf{X}_{odd}\right) \tag{3}$$

$$\mathbf{D}^b = g_{AAE2}\left(\mathbf{X}_{even}\right) \tag{4}$$

where $\mathbf{D}^a = \left[\mathbf{d}_1^a, \mathbf{d}_2^a, \ldots, \mathbf{d}_{H\times W}^a\right] \in \mathbb{R}^{(H\times W)\times L}$ and $\mathbf{D}^b = \left[\mathbf{d}_1^b, \mathbf{d}_2^b, \ldots, \mathbf{d}_{H\times W}^b\right] \in \mathbb{R}^{(H\times W)\times L}$ are the final data augmentation samples.

2.2. Contrastive head

The contrastive head includes two parts, named as spectrum contrastive head and cluster contrastive head. The spectrum contrastive head is a two-layer non-linear multilayer perceptron (MLP), denoted as $g_S(\cdot)$, where the number of neurons in each layer is the same. The cluster contrastive head is also a MLP with two layers, denoted as $g_c(\cdot)$, where the number of neurons in the second layer represents the number of clusters. To separate target and background, set the number of clusters to 2. In contrastive learning, formally, N augmented samples are taken from the same position in $\mathbf{D}^a \in \mathbb{R}^{(H\times W)\times L}$ and $\mathbf{D}^b \in \mathbb{R}^{(H\times W)\times L}$ respectively, expressed as $\mathbf{D}_{batch}^a = \left[\mathbf{d}_1^a, \ldots, \mathbf{d}_N^a\right] \in \mathbb{R}^{N\times L}$ and $\mathbf{D}_{batch}^b = \left[\mathbf{d}_1^b, \ldots, \mathbf{d}_N^b\right] \in \mathbb{R}^{N\times L}$. Then the backbone is used to extract representation vectors of $\mathbf{D}_{batch}^a$ and $\mathbf{D}_{batch}^b$, which can

be expressed as $\mathbf{H}^a = \left[\mathbf{h}_1^a, \ldots, \mathbf{h}_N^a\right] \in \mathbb{R}^{N \times L/4}$ and $\mathbf{H}^b = \left[\mathbf{h}_1^b, \ldots, \mathbf{h}_N^b\right] \in \mathbb{R}^{N \times L/4}$. The backbone is a deep residual convolutional neural network with spectral residual channel attention, which is used to extract representation vectors from augmented data samples. For the spectrum contrastive head, the representation vector $\mathbf{h}_i^a$ is paired with another augmented sample of the pixel spectrum at the same position through the representation $\mathbf{h}_i^b$ extracted by the backbone to form a positive pair $\left\{\mathbf{h}_i^a, \mathbf{h}_i^b\right\}$, and the remaining representation vectors form negative pairs. Then the spectrum contrastive head $g_S(\cdot)$ is used to map the representations extracted by the backbone to the spectral contrast space via $\mathbf{z}_i^a = g_s\left(\mathbf{h}_i^a\right)$ and $\mathbf{z}_i^b = g_s\left(\mathbf{h}_i^b\right)$. The similarity between the pairs is measured by cosine distance, which can be expressed as

$$s\left(\mathbf{z}_i^{c_1}, \mathbf{z}_j^{c_2}\right) = \frac{\left(\mathbf{z}_i^{c_1}\right)\left(\mathbf{z}_j^{c_2}\right)^{\mathrm{T}}}{\left\|\mathbf{z}_i^{c_1}\right\|\left\|\mathbf{z}_j^{c_2}\right\|} \tag{5}$$

where $c_1, c_2 \in \{a, b\}$ and $i, j \in [1, N]$. The spectral contrast loss for the representation vector $\mathbf{h}_i^a$ can be defined as

$$l_i^a = -\log \frac{\exp\left(s\left(\mathbf{z}_i^a, \mathbf{z}_i^b\right)/\tau_s\right)}{\sum_{j=1}^{N}\left[\exp\left(s\left(\mathbf{z}_i^a, \mathbf{z}_j^a\right)/\tau_s\right) + \exp\left(s\left(\mathbf{z}_i^a, \mathbf{z}_j^b\right)/\tau_s\right)\right]} \tag{6}$$

where τ_s is the spectral temperature parameter to control the softness. The spectral contrast loss is calculated for each representation vector, and the loss function of the spectrum contrastive head is obtained as

$$L_{spectral} = \frac{1}{2N}\sum_{i=1}^{N}\left(l_i^a + l_i^b\right) \tag{7}$$

For the cluster contrastive head, the cluster contrastive head $g_c(\cdot)$ is used to map the representations extracted by the backbone to cluster contrast space via $\mathbf{Y}^a = g_c\left(\mathbf{H}^a\right)$ and $\mathbf{Y}^b = g_c\left(\mathbf{H}^b\right)$. The dimension of the row vectors $\mathbf{y}_i^a$ and $\mathbf{y}_i^b$ in the feature matrices $\mathbf{Y}^a = \left[\mathbf{y}_1^a, \ldots, \mathbf{y}_N^a\right] \in \mathbb{R}^{N \times 2}$ and $\mathbf{Y}^b = \left[\mathbf{y}_1^b, \ldots, \mathbf{y}_N^b\right] \in \mathbb{R}^{N \times 2}$ are both 2. It can be considered that the *j*th element in the row vectors $\mathbf{y}_i^a$ and $\mathbf{y}_i^b$ represents the probability that the sample belongs to the *j*th cluster. As a result, from the perspective of feature matrix columns, let $\tilde{\mathbf{y}}_i^a$ and $\tilde{\mathbf{y}}_i^b$ represent the *i*th column in feature matrixes $\mathbf{Y}^a$ and $\mathbf{Y}^b$, respectively. Then the $\tilde{\mathbf{y}}_i^a$ is paired with $\tilde{\mathbf{y}}_i^b$ to positive cluster pair $\left\{\tilde{\mathbf{y}}_i^a, \tilde{\mathbf{y}}_i^b\right\}$, while leaving other 2 pairs to be negative. To distinguish cluster $\tilde{\mathbf{y}}_i^a$ from all other clusters except $\tilde{\mathbf{y}}_i^b$, the cluster contrast loss is defined as

$$\tilde{l}_i^a = -\log \frac{\exp\left(s\left(\tilde{\mathbf{y}}_i^a, \tilde{\mathbf{y}}_i^b\right)/\tau_c\right)}{\sum_{j=1}^{2}\left[\exp\left(s\left(\tilde{\mathbf{y}}_i^a, \tilde{\mathbf{y}}_j^a\right)/\tau_c\right) + \exp\left(s\left(\tilde{\mathbf{y}}_i^a, \tilde{\mathbf{y}}_j^b\right)/\tau_c\right)\right]} \tag{8}$$

where τ_c is the cluster temperature parameter to control the softness. By traversing the target and background clusters, the loss function of the cluster contrastive head is

$$L_{cluster} = \frac{1}{2 \times 2}\sum_{i=1}^{2}\left(\tilde{l}_i^a + \tilde{l}_i^b\right) \tag{9}$$

2.3. Objective function

The spectrum contrastive head and the cluster contrastive head are simultaneously optimized, so that the backbone has the ability to distinguish between spectral similarity and dissimilarity. The objective function of the contrastive learning stage is

$$L = L_{spectral} + L_{cluster} \tag{10}$$

After contrastive learning, the backbone is used to extract the representation vector of each pixel spectrum and the prior target spectrum in the HIS to be detected. Then the cosine distance is used to measure the similarity between the pixel spectrum in the HSI to be detected and the prior target spectrum, and the detection result is obtained.

3. EXPERIMENT

To evaluate the performance of the proposed CLHTD detector, several detection methods, including ACE, CEM, CSCR [6], ECEM [7] and CNNTD [1] are applied for comparison. The dataset is part of San Diego airport, California, USA. The spatial size is 120×120 . After removing low SNR and water absorption bands, 189 bands are reserved for hyperspectral target detection.

Fig. 2 (a) and (b) show the pseudo color image and ground truth of San Diego dataset. For the proposed CLHTD method, data augmentation is first by sampling the odd and even bands, and then the two types of augmentation are obtained by the encoder in the trained AAE. When training AAE, the encoder and decoder are optimized by Adam optimizer, and the learning rate is set to 0.001. Then the generator and discriminator are trained, and the learning rate is set to 0.0001 and 0.00001 when SGD optimizer is used to optimize the generator and discriminator. The batch size and epoch when training the AAE are set to 240 and 20. In contrastive learning, the epoch, batch size, and learning rate are set to 100, 240, 0.05. The temperature parameter τ_s and τ_c are both set to 0.1. Fig. 2 (c) to (h) show the detection maps of different methods on the San Diego dataset. It can be observed that our CLHTD significantly highlights the target

and suppresses the background compared to other detection methods.

To quantitatively analyze the detection algorithm, the ROC curve of (P_D, P_F) and ROC curve of (P_F, τ) are used to assess the detection and background suppression performance, as shown in Fig. 3 (a) and Fig. 3 (b), and the corresponding AUC value is shown in Table 1.

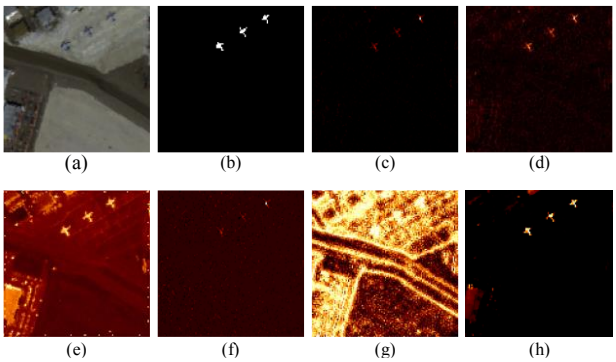
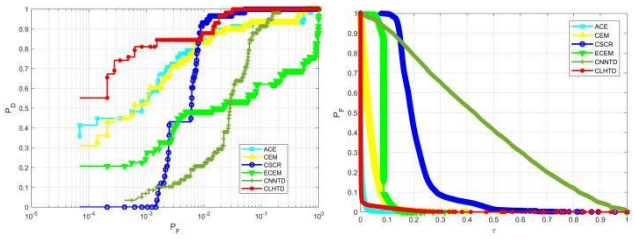


Fig. 2. San Diego dataset and detection maps. (a) Pseudo color image, (b) ground truth, (c) ACE, (d) CEM, (e) CSCR, (f) ECEM, (g) CNNTD, (h) CLHTD

By observing Fig. 3 (a) and Fig. 3 (b), it is obvious that the CLHTD showed superior detection and background suppression ability. Fig. 4 shows the separability graphs of detection results for six detection methods of San Diego dataset. The CLHTD can better separate the target from the background.



(a) ROC curve of (P_D, P_F) (b) ROC curve of (P_F, τ)

Fig. 3. ROC curves of different algorithms on San Diego dataset.

Table 1. Accuracy comparison of different methods.

Method	ACE	CEM	CSCR	ECEM	CNNTD	CLHTD
$AUC_{(P_D,P_F)}$	0.9558	0.9628	0.9936	0.7049	0.9580	**0.9972**
$AUC_{(P_F,\tau)}$	**0.0042**	0.0385	0.2112	0.0870	0.4481	0.0058

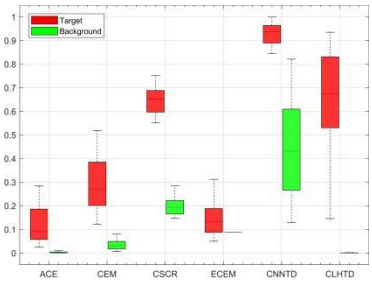


Fig. 4. Separability graphs of detection results for six test methods based on San Diego dataset.

4. CONCLUSION

In this work, a hyperspectral target detection method based on contrastive learning is proposed. The positive and negative pairs are constructed through data augmentation. In the contrastive learning phase, the backbone is used to extract the representation vector of the positive and negative pairs. Then, the representation vectors are mapped to the spectral contrast space by the spectrum contrastive head to learn spectral similarity and dissimilarity. At the same time, the cluster contrastive head maps the representation vector to the cluster contrast space to cluster them into two classes to increase the separation of the target and background. The backbone with the ability to distinguish spectral similarity and dissimilarity can then extract the representation vector of the prior target spectrum and the pixel spectrum in the HSI to be detected, and then measure the similarity by cosine distance to obtain target detection results. The experiments show that our CLHTD method is superior to other comparison detectors.

5. ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of China (61801075, 42101350), China Postdoctoral Science Foundation (2020M670723) and the Fundamental Research Funds for the Central Universities (3132022232).

REFERENCES

[1] W. Li, G. Wu, and Q. Du, "Transferred deep learning for hyperspectral target detection," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 5177-5180.

[2] 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 Sensing,* vol. 12, no. 9, 2020.

[3] W. Xie, X. Zhang, Y. Li, K. Wang, and Q. Du, "Background Learning Based on Target Suppression Constraint for Hyperspectral Target Detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 13, pp. 5887-5897, 2020.

[4] W. Xie, J. Yang, J. Lei, Y. Li, Q. Du, and G. He, "SRUN: Spectral Regularized Unsupervised Networks for Hyperspectral Target Detection," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 58, no. 2, pp. 1463-1474, 2020.

[5] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

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

[7] R. Zhao, Z. Shi, Z. Zou, and Z. Zhang, "Ensemble-Based Cascaded Constrained Energy Minimization for Hyperspectral Target Detection," *Remote Sensing,* vol. 11, no. 11, 2019.