

Feature Enhancement with Reverse Distillation for Hyperspectral Anomaly Detection

Wenping Jin, Feng Dang, Li Zhu

Abstract—Distinguished from most hyperspectral Anomaly Detection(AD) methods based on trainable parameter networks, the recently proposed method called AETNet eliminates the need for parameter adjustments or retraining on new test scenes by training an anomaly enhancement network on background data with false anomalies. In this letter, we achieve this by proposing a novel training and inference framework that enhances the network’s background spectral feature extraction capability without any data augmentation. During training on background data, the complete network is trained using the reverse distillation framework with a spectral feature alignment mechanism to improve the network’s background feature expressiveness. For inference, a pruned network is applied, composed solely of components most relevant to expressing features in the spectral dimension. This effectively reduces redundant information, enhancing both inference efficiency and anomaly detection accuracy. Experimental results demonstrate that our method outperforms state-of-the-art methods on the HAD100 dataset, striking an optimal balance between detection accuracy and inference speed. Our code is available at <https://github.com/cristianoKaKa/FERD>.

Index Terms—Anomaly detection, hyperspectral image (HSI), reverse distillation.

I. INTRODUCTION

HYPERSPECTRAL remote sensing technology provides a rich tapestry of spectral information by capturing a broad range of wavelengths from the visible to the infrared [1], [2]. This depth of spectral data opens up expansive applications, notably in agriculture [3], resource exploration [4], environmental monitoring [5], and military reconnaissance [6]. Anomaly Detection(AD) is key in hyperspectral image (HSI) analysis, focusing on identifying pixels whose spectral signatures starkly contrast with their immediate environment [7], [8]. By highlighting such spectral anomalies, one can detect key indicators of geological variations, landscape changes, shifts in biodiversity, alterations in environmental conditions, or evidence of human activities—each of which is of critical interest within their respective application fields. The Reed-Xiaoli (RX) algorithm stands as a fundamental method in hyperspectral AD [9]. It is premised on the assumption that the background conforms to a Gaussian distribution. It calculates the Mahalanobis distance to identify anomalous pixels that exhibit spectral properties diverging from those expected under the Gaussian model. RX operates in two variants: the global RX (GRX), which utilizes a global background model, and the local RX (LRX), which uses a local background spectrum [10].

Wenping Jin, Feng Dang, Li Zhu are with School of Software Engineering, Xi’an Jiaotong University, Xi’an 710049, China. (e-mails: jinwenping@stu.xjtu.edu.cn; 3121358205@stu.xjtu.edu.cn; zhuli@xjtu.edu.cn). Li Zhu is the corresponding author.

While the RX algorithm has achieved widespread acclaim in the realm of hyperspectral AD, its limitations are notable, particularly with non-Gaussian data distributions and in noisy environments. To overcome these challenges, recent developments have witnessed a surge in trainable parameter network methods, such as autonomous hyperspectral AD network (AutoAD) [11], low-rank embedded network (LREN) [12], and convolutional transformer-inspired autoencoder (CTA) [13]. These innovative approaches aim to capture complex data representations more adeptly, thus enhancing AD capabilities. Nonetheless, a significant drawback of these methods is their need for extensive retraining when applied to new HSIs containing anomalies, leading to considerable time expenditure. To address these issues, Zhaoxu Li et al. introduced Anomaly Enhancement Transformation (AETNet) [14]. AETNet trains a denoising reconstruction network using anomaly-free HSIs with false anomalies, enabling the reconstruction network to enhance anomaly features in HSIs. During the inference phase, it no longer undergoes training; instead, it utilizes the reconstruction network to enhance the original HSI data features, followed by the application of non-parametric AD algorithms such as RX to detect anomalous pixels. This method has achieved higher accuracy compared to non-parametric AD algorithms on the extensive HAD100 hyperspectral AD dataset [14], almost without compromising speed. This holds great practical significance in hyperspectral AD. Our research is devoted to refining this method of hyperspectral AD, aiming to further increase detection accuracy and accelerate inference, meanwhile tackling some of AETNet’s limitations such as its heavy reliance on data augmentation techniques.

We employ the concept of Reverse Distillation (RD) [15] during training, without resorting to any data augmentation techniques. Unlike traditional Autoencoders, RD treats the encoder as a ‘teacher’ network and the decoder as a ‘student’ network. The teacher encoder is typically a well-pretrained network with fixed parameters, and the student decoder learns the inverse representation from the teacher. Thanks to its asymmetric structure, RD effectively addresses the overfitting and underfitting issues encountered by Autoencoders in feature extraction. In conventional RD methods, regions in the latent space that the decoder cannot effectively reconstruct are often regarded as anomaly areas. However, directly applying RD to the hyperspectral AD domain is not advisable due to its emphasis on spatial dimension information while neglecting spectral dimension information. To address this, in the training phase, we introduce the Spectral Feature Alignment Mechanism (SFAM), which extracts lower-level features from both the teacher encoder and student decoder that are more

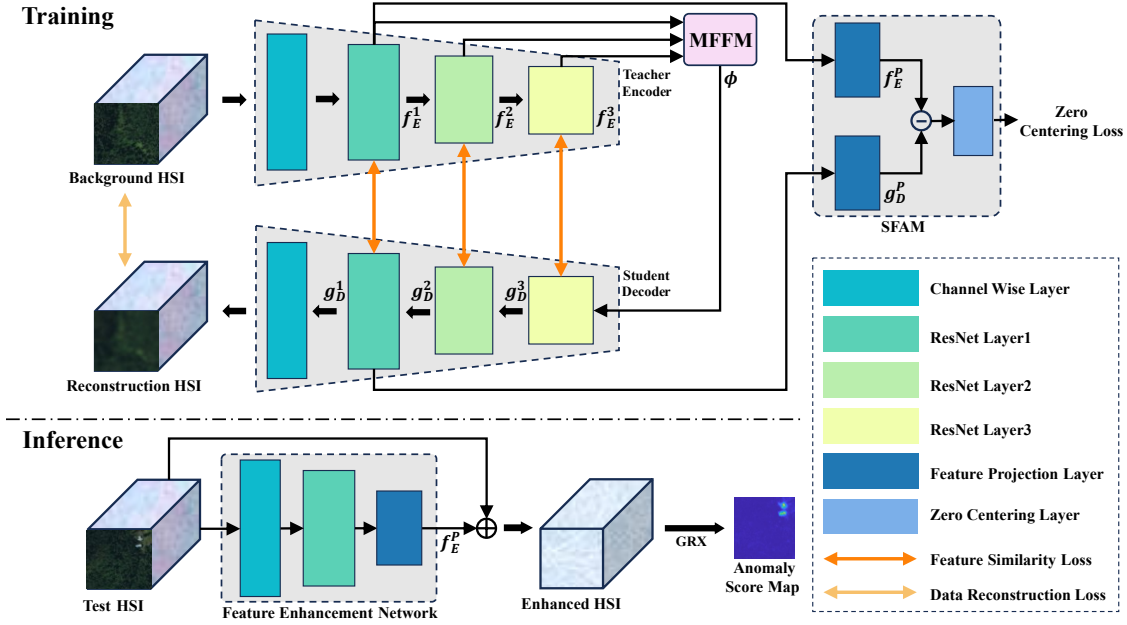


Fig. 1. Proposed method overview.

relevant to spectral information. SFAM aligns these features in the original spectral dimension. During inference, we utilize a pruned network structure composed solely of components most relevant to expressing features in the spectral dimension as a Feature Enhancement Network (FEN) to enhance the original HSI data features rather than directly using reconstruction error as the basis for anomaly detection. Finally, we employ non-parametric AD algorithms such as GRX based on the enhanced HSI to detect anomalies. Our method improves the model's ability to extract spectral features, reduces interference from redundant information, enhances detection accuracy, and reduces inference time. Ultimately, through validation on the HAD100 dataset, our approach has demonstrated superior performance and faster inference compared to AETNet.

II. PROPOSED METHOD

The overall training and inference procedure of the proposed method is illustrated in Fig. 1. In this section, we will provide a detailed description of this process.

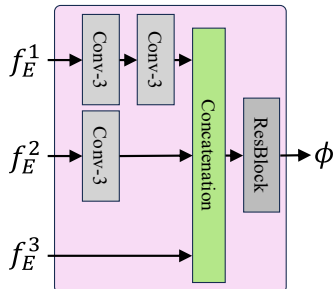


Fig. 2. Multi-scale feature fusion module.

A. Training

During the training phase, the complete network architecture is utilized, comprising the teacher encoder, Multi-scale Feature Fusion Module (MFFM), the student decoder and SFAM.

The teacher encoder includes the Channel Wise Layer (CWL) and layers 1, 2, and 3 from the ResNet. The CWL of the encoder is a 1x1 Convolutional Neural Network (CNN) layer with input dimensions equal to the spectral dimensions of input background HSI and output dimensions equal to input channel of ResNet layer 1, replacing the original initial CNN layer in the ResNet network. One crucial point to emphasize here is that, in the classical RD training process, the teacher encoder should be a well-pretrained network with fixed parameters. However, in our method, as the CWL serves as the first layer of the encoder, responsible for extracting the fundamental spectral features of the input HSI, it needs to be adjusted during training to ensure it can effectively capture these basic features. In contrast, the other layers in the encoder deal with more abstract features. Therefore, we directly employ generic pretrained parameters for these layers and keep them fixed throughout training. Let us denote by H an input HSI, and the outputs of layers 1, 2, and 3 of ResNet are multi-scale features f_E^1 , f_E^2 , and f_E^3 from low to high levels, the encoder process can be represented as: $f_E^1 = E^1(C_E(H))$, $f_E^2 = E^2(f_E^1)$, and $f_E^3 = E^3(f_E^2)$. Where E^k represent the layer k of ResNet in encoder, C_E is the encoder's CWL.

The MFFM is trainable, and its role is to generate fused features ϕ from the multi-scale features of the encoder, which serve as inputs to the decoder, as shown in Fig. 2.

The student decoder structure mirrors the encoder, utilizing the ResNet as the backbone but with convolutions replaced by deconvolutions. The multi-scale output features of the decoder are denoted as g_D^3 , g_D^2 , and g_D^1 , expressed as: $g_D^3 = D^3(\phi)$,

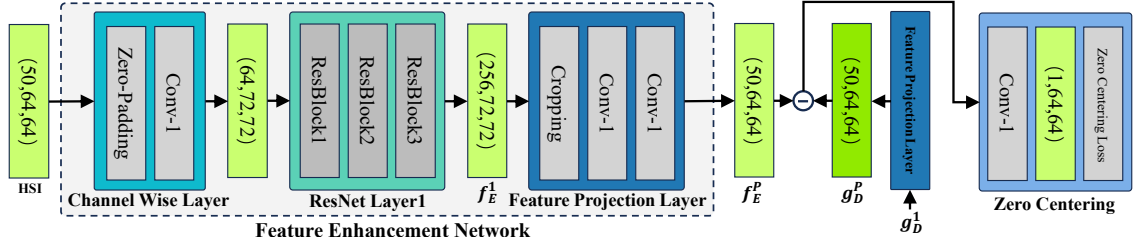


Fig. 3. Detailed process of spectral feature alignment mechanism.

$g_D^2 = D^2(g_D^3)$, and $g_D^1 = D^1(g_D^2)$, where D^k represents the layer k of the ResNet in the decoder. Similarly, the decoder employs its CWL to transform the output features into the reconstructed HSI, denoted as $H_R = C_D(g_D^1)$. The decoder learns the inverse representation of the encoder through Feature Reconstruction Loss \mathcal{L}_F and Data Reconstruction Loss \mathcal{L}_D , computed as follows:

$$\mathcal{L}_F = 1 - \frac{1}{N} \sum_{i=1}^N \text{CosSim}(g_D^i, f_E^i)$$

$$\mathcal{L}_D = \text{MSE}(H, H_R).$$

Here, CosSim represents the cosine similarity loss function [16], and MSE stands for the mean squared error loss function.

The SFAM consists of a Feature Projection Layer (FPL) and a Zero Centering Layer (ZCL). The FPL comprises multiple 1×1 CNN layers, responsible for pixel-wise mapping of low-level features f_E^1 and g_D^1 from the encoder and decoder to the spectral space of HSI, denoted as f_E^P and g_D^P respectively. Fig. 3 illustrates the detailed processing of the input HSI with spatial size 64×64 and spectral dimension 50 through SFAM. As shown in the figure, from the input to f_E^P , this process constitutes the construction of the FEN, with all layers except for the three 3×3 CNN layers within the ResBlocks being 1×1 CNN layers. This design ensures that f_E^P primarily contains spectral features while preserving neighboring spatial features. On the other hand, g_D^P represents the feature projection from the decoder, incorporating advanced features that integrate multi-scale spatial information. However, due to the complexity of the RD process, g_D^P also contains noise introduced by incomplete optimization of the decoder. We aim for the FEN to learn advanced feature representations while remaining unaffected by noise. Thus, we conservatively align f_E^P and g_D^P by projecting $|f_E^P - g_D^P|$ onto a one-dimensional space as O using ZCL, which is a single 1×1 CNN layer, and optimizing O towards a mean of 0. This process's loss is referred to as Zero Centering Loss and denoted as \mathcal{L}_Z , defined as follows:

$$\mathcal{L}_Z(O) = -\frac{1}{N} \sum_{i=1}^N \log\left(1 - \frac{1}{1 + e^{-o_i}}\right),$$

where o_i represents the i -th element of matrix O .

The final loss function for the training process is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_F + \lambda_2 \mathcal{L}_D + \lambda_3 \mathcal{L}_Z,$$

where λ_1 , λ_2 , and λ_3 are the weights for each loss component.

B. Inference

In inference, we directly utilize the output of the FEN as enhanced features rather than the reconstruction error of the complete network. This decision stems from two main factors. Firstly, during training, the output of the FEN already possesses advanced feature representation capabilities. Secondly, using the reconstruction error would inevitably introduce unnecessary noise resulting from the incomplete optimization of the decoder. Given a test HSI H , to compensate for the loss of some original features caused by the processing through the FEN, its enhanced version is defined as the sum of H and the output of the FEN, $H_E = H + \text{FEN}(H)$. The anomaly score of each spectrum h_E^i on H_E is calculated using the GRX algorithm.

III. EXPERIMENTS

A. Datasets and Evaluation Metrics

Empirical evaluations were conducted on the HAD100 hyperspectral anomaly detection dataset, which consists of 100 authentic remote sensing test scenes. These scenes include marked targets, predominantly compact manufactured objects such as vehicles, boats, and buildings, with sizes ranging from 1 pixel to 69 pixels. Additionally, they encompass various backgrounds such as grasslands, forests, farmlands, deserts, lakes, rivers, and coastlines. The HAD100 dataset acquired all test hyperspectral images (HSIs) from the AVIRIS-NG website and uniformly cropped them into blocks of size 64×64 . Furthermore, HAD100 provides two training sets captured separately by AVIRIS-NG and AVIRIS-Classic. The HSIs in these training sets are cropped only from background regions. In our experiments, we utilized background data training sets captured by AVIRIS-NG and employed 100 authentic remote sensing test scenes as the testing set.

The evaluation metric adopted is the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which is the most widely used evaluation tool in hyperspectral AD tasks. The effectiveness of the methods is evaluated on the HAD100 dataset using the mean AUC value (mAUC) across all 100 HSIs in the testing set.

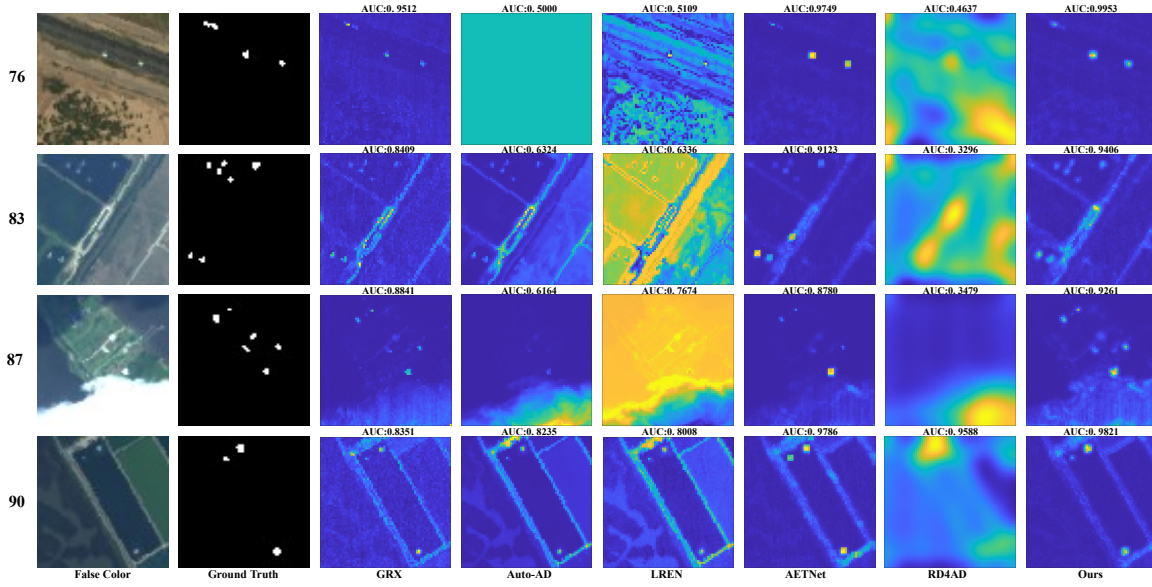


Fig. 4. Detection maps of different AD methods on the first 50 bands of the test scenes 76, 83, 87, and 90.

B. Implementation Details

In the experiment, we compared our proposed method with several state-of-the-art methods, including the global RX based on nonparametric estimation (GRX), Auto-AD and LREN based on trainable parameter network, AETNet based on anomaly enhancement network, and the original RD method. The code for GRX is a widely used unofficial version, while the remaining comparative methods use official code. To adapt the original RD method to HSI, the first layer of the encoder and the last layer of the decoder are replaced with CWL, while the rest remain unchanged from the official code. All codes were used to compute the AUC for 100 test samples in HAD100.

The parameter settings for our method during training are as follows: we utilized the Adam optimization algorithm with a fixed learning rate of 0.005, momentum values set to $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The batch size was set to 16, and the networks were trained for 60 epochs. The coefficients for the loss terms were set to $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$. Test results were obtained from the best-performing model over the 60 epochs. Additionally, our experiments were conducted on a computer equipped with Intel i7 11700T CPU (16M Cache, up to 4.60 GHz), 64GB of RAM, and Nvidia GeForce RTX 3090Ti GPU.

C. Experimental Results

In Table I, we provide a comprehensive performance comparison of various AD methods using the HAD100 test dataset. The evaluation is based on two key performance metrics: the mAUC and the average computation time in seconds. Our proposed method, denoted as "Ours," consistently outperforms existing methods across different spectral bands of the HAD100 dataset in terms of mAUC. Specifically, for the first 50 bands, OURS achieves the highest mAUC of 0.9941, showcasing superior detection accuracy. This trend continues

for the first 100 and 200 bands, with mAUC values of 0.9901 and 0.9890, respectively. Importantly, OURS demonstrates consistently low computation times, being only around 0.02 seconds slower than GRX and approximately 0.005 seconds faster than AETNet, while maintaining superior detection accuracy compared to AETNet. In contrast, methods based on trainable parameter networks like Auto-AD and LREN require extensive retraining on each test scene, leading to longer computation times and inferior accuracy compared to traditional methods like GRX. While the original RD algorithm (denoted as "RD4AD") performs well in conventional image AD, its effectiveness is limited when applied to hyperspectral AD.

Fig. 4 illustrates the detection maps generated by various AD methods on the first 50 bands of the HAD100 test set. Specifically, we focus on the most challenging test scenes numbered 76, 83, 87, and 90. It is evident from the visual results that our method achieves the highest detection accuracy in these scenes, as it produces clear target outlines while effectively suppressing background noise.

TABLE I
PERFORMANCE COMPARISON OF METHODS ON HAD100

Method	The First 50 Bands		The First 100 Bands		The First 200 Bands	
	mAUC	Time	mAUC	Time	mAUC	Time
GRX	0.9799	0.017	0.9714	0.024	0.9649	0.052
Auto-AD	0.8636	7.236	0.7040	7.412	0.6825	7.527
LREN	0.8858	33.437	0.8820	35.389	0.8771	38.523
AETNet	0.9925	0.043	0.9875	0.049	0.9818	0.078
RD4AD	0.8053	0.027	0.7542	0.027	0.8196	0.028
Ours	0.9941	0.037	0.9901	0.044	0.9890	0.073

D. Ablation Study

Table II presents the mAUC values on the first 50 bands of the HAD100 dataset under different combinations of loss functions. Each loss activation indicates whether the corresponding

structure is enabled during model training: \mathcal{L}_Z corresponds to the SFAM, \mathcal{L}_D and \mathcal{L}_F respectively indicate whether input and multi-scale features reconstruction occurs in the student decoder. As depicted in the table, activating any single loss function yields the poorest performance. However, when any two loss functions are activated, there is a noticeable improvement. Notably, when all three loss functions (\mathcal{L}_Z , \mathcal{L}_D , and \mathcal{L}_F) are simultaneously enabled, a significant enhancement in detection performance is observed. This underscores the synergistic effect of integrating the SFAM, input reconstruction, and multi-scale feature reconstruction in improving the accuracy of AD.

TABLE II
DIFFERENT LOSS FUNCTIONS ON AD PERFORMANCE

\mathcal{L}_Z	\mathcal{L}_D	\mathcal{L}_F	Mean AUC
✓			0.9913
	✓		0.9915
		✓	0.9924
✓	✓		0.9917
✓		✓	0.9924
	✓	✓	0.9934
✓	✓	✓	0.9941

Table III demonstrates the performance of AD with various feature enhancement methods. Here, H represents the original HSI, f_E^P denotes the output of the FEN, and g_D^P signifies the projection of low-level features from the decoder in SFAM. As observed, consistent with our analysis in the section II, utilizing the original HSI combined with the output of FEN as enhanced features yields optimal results.

TABLE III
DIFFERENT FEATURE ENHANCEMENT METHODS ON AD PERFORMANCE

Input of GRX	mAUC/Time
H	0.9799/0.017
f_E^P	0.9934/0.037
$H + f_E^P$	0.9941/0.037
$H + g_D^P$	0.9689/0.181
$H + f_E^P - g_D^P $	0.9935/0.181

IV. CONCLUSION

In this letter, we present a novel method for hyperspectral AD by employing the training framework of RD with SFAM and unique inference based on the pruned network. We demonstrate the outstanding performance of our method on the large-scale hyperspectral AD dataset HAD100. Additionally, the rationality of our method is validated through the ablation study.

However, it is worth noting that our training process is cumbersome, and most structures are discarded during inference. Future research could focus on exploring more compact and efficient models and training methods.

V. ACKNOWLEDGMENTS

The authors thank National Key Research and Development Program of China (2021QY0801) for funding support.

REFERENCES

- [1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi and J. Chanussot, "Hyperspectral Remote Sensing Data Analysis and Future Challenges," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6-36, June 2013, doi: 10.1109/MGRS.2013.2244672.
- [2] Goetz, A. Three decades of hyperspectral remote sensing of the Earth: A personal view. *Remote Sensing Of Environment*. **113** pp. S5-S16 (2009), <https://www.sciencedirect.com/science/article/pii/S003442570900073X>, Imaging Spectroscopy Special Issue.
- [3] N. K. Patel & Dave, A. Study of crop growth parameters using Airborne Imaging Spectrometer data. *International Journal Of Remote Sensing*. **22**, 2401-2411 (2001).
- [4] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," in *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 12-16, Jan. 2002, doi: 10.1109/79.974715.
- [5] Govender, M., Chetty, K. & Bulcock, H. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa*. **33**, 145-151 (2007).
- [6] S. Khazai, A. Safari, B. Mojaradi and S. Homayouni, "An Approach for Subpixel Anomaly Detection in Hyperspectral Images," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 769-778, April 2013, doi: 10.1109/JSTARS.2012.2210277.
- [7] H. Su, Z. Wu, H. Zhang and Q. Du, "Hyperspectral Anomaly Detection: A survey," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 64-90, March 2022, doi: 10.1109/MGRS.2021.3105440.
- [8] Racetin, I. & Krtalić, A. Systematic Review of Anomaly Detection in Hyperspectral Remote Sensing Applications. *Applied Sciences*. **11** (2021), <https://www.mdpi.com/2076-3417/11/11/4878>.
- [9] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1760-1770, Oct. 1990, doi: 10.1109/29.60107.
- [10] J. M. Molero, E. M. Garzón, I. García and A. Plaza, "Analysis and Optimizations of Global and Local Versions of the RX Algorithm for Anomaly Detection in Hyperspectral Data," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 801-814, April 2013, doi: 10.1109/JSTARS.2013.2238609.
- [11] S. Wang, X. Wang, L. Zhang and Y. Zhong, "Auto-AD: Autonomous Hyperspectral Anomaly Detection Network Based on Fully Convolutional Autoencoder," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022, Art no. 5503314, doi: 10.1109/TGRS.2021.3057721.
- [12] Jiang, K., Xie, W., Lei, J., Jiang, T. & Li, Y. LREN: Low-rank embedded network for sample-free hyperspectral anomaly detection. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **35**, 4139-4146 (2021).
- [13] Z. He, D. He, M. Xiao, A. Lou and G. Lai, "Convolutional Transformer-Inspired Autoencoder for Hyperspectral Anomaly Detection," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023, Art no. 5508905, doi: 10.1109/LGRS.2023.3312589.
- [14] 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," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-18, 2023, Art no. 5506718, doi: 10.1109/TGRS.2023.3258067.
- [15] Deng, H. & Li, X. Anomaly Detection via Reverse Distillation From One-Class Embedding. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 9737-9746 (2022,6).
- [16] Schroff, F., Kalenichenko, D. & Philbin, J. Deep metric learning using triplet network. *ArXiv Preprint ArXiv:1412.6622*. (2015).