# A CONTRASTIVE LEARNING METHOD FOR MULTI-LABEL PREDICTORS ON HYPERSPECTRAL IMAGES

*Salma Haidar*

University of Antwerp, imec-IDLab, MicroTechniX BV, Belgium

*José Oramas*

University of Antwerp imec-IDLab, Belgium

## ABSTRACT

Self-supervised contrastive learning is increasingly acknowledged as an effective approach to mitigate the challenges posed by limited annotated data. We introduce a two-stage methodology that extends current approaches, targeting the downstream task of multi-label classification in hyperspectral remote-sensing imagery. In the initial stage, we employ a contrastive learning approach to train a base encoder and a projection neural network, thereby learning data patterns without relying on annotations. The effectiveness of the encoder is bolstered as it is guided by a contrastive loss function to maximize the similarity between the generated embeddings. In the second stage, we harness the power of the pre-trained encoder to channel its hidden representations into a multi-label classifier. Our empirical validation demonstrate that this method surpasses fully supervised alternatives. The observed improvements are attributed to the strategy of training the encoder alongside the classifier, thereby refining its adaptability to the feature space of the classifier.

***Index Terms*—** Hyperspectral imagery, remote sensing, self-supervised learning, contrastive learning, multi-label classification, deep learning

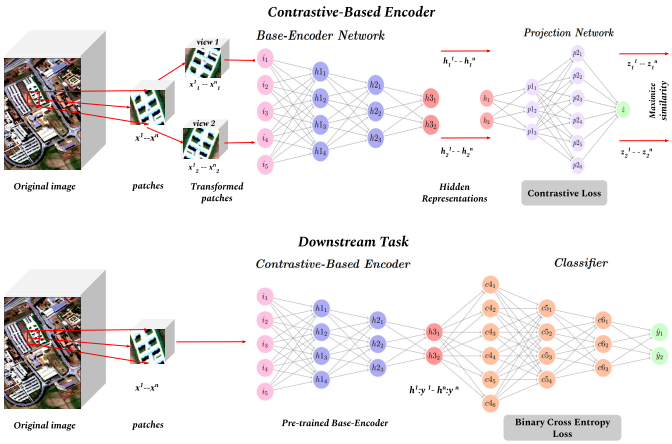


**Fig. 1**. Two-stage method: Stage one trains an encoder in a contrastive manner on unlabelled data, maximizing agreement between hidden representations of two transformations from the same patch sample. Stage two fine-tunes the pre-trained encoder with a classifier on labelled samples for multi-label classification

## 1. INTRODUCTION

The generalisation capacity of deep learning methods relies heavily on the availability of large, carefully labelled datasets. However, manual annotation of data is an expensive, time-consuming, and tedious process that often lacks domain expertise, leading to potential issues with annotation quality. These limitations have motivated research in the domain of self-supervised learning (SSL) [1, 2]. It leverages the learning of useful representations from unlabelled data depending solely on the intrinsic patterns in the data. Those representations are then used in downstream tasks that usually require large sets of labelled data for a successful training. Unlike general unsupervised learning methods, which focus on discovering hidden patterns in unlabelled data without explicit guidance, self-supervised learning methods take a different approach. They create pretext tasks derived from the data itself to generate labels. This allows the models to learn more targeted and task-specific representations.

Contrastive learning (CL) has a long standing history [3, 4, 5], but its recent application in self-supervised representation learning for computer vision [6] has drawn significant attention. It is a technique that compares input data with similar or contrasting data to derive representations for a specific task. A crucial aspect of CL is the design of a contrastive loss function [7] that encourages the algorithm to learn corresponding representations for similar inputs and distinct representations for dissimilar inputs. When applied in a self-supervised manner, CL operates on unlabelled datasets to uncover and learn representations of the data based on its inherent characteristics. The loss function encourages the model to learn representations capable of distinguishing between two perspectives of the same input sample. In self-supervised contrastive learning, a pretext task is employed, involving the generation of pairs of augmented views of the same data sample (positive pairs) and contrasting them with views of different samples (negative pairs), thereby facilitating the learning of insightful representations from the data. The learned weights can then be transferred to the final intended task.

Images, typically in RGB format, have been at the heart of the self-supervised CL research in the literature. However, one interesting avenue of research would be to extend this application to hyperspectral remote sensing images characterised by a deep spectral extent. Hyperspectral imagery captures the reflectance of objects across contiguous wavelength bands of the electromagnetic spectrum resulting in a hypercube with both spatial and spectral dimensions. Each pixel in this imagery represents a unique combination of colour and wavelength. In the field of remote sensing, those images are used to map the distribution of land cover categories, such as forest, crops and urban areas. Deep learning has shown considerable promise in the field of hyperspectral image (HSI) analysis for remote sensing applications, particularly over the last decade [8, 9, 10]. This is because deep models such as Convolutional Neural Networks (CNNS) and Recurrent Neural Networks (RNNs) have shown exceptional ability to automatically learn relevant features from high-dimensional data. Hyperspectral images possess both high spatial

and spectral dimensionality, offering an abundance of information that can benefit deep learning algorithms. However, this characteristic also brings forth complex challenges for deep learning methodologies designed for HSI analysis including computational complexity and the "curse of dimensionality," which can adversely affect model training and performance [11, 12, 13]. These challenges arise from several factors such as: 1) the scarcity of labelled data, 2) the high sensitivity to network architectures, and 3) the employed training schemes.

Notwithstanding the aforementioned factors self-supervised CL emerges as a viable alternative to supervised deep learning methods in the domain of hyperspectral imagery analysis. In this paper, we propose a relatively deep, yet simple architecture of an encoder (feature extractor) and a classifier. We employ a self-supervised contrastive learning method to train the encoder to provide high-quality data representations to the classifier. With this enhanced input, the classifier is better equipped to accurately make multi-label predictions for patches of hyperspectral remote sensing images. We adopt the concept followed in [14] while using a different architecture and focusing on geometric transformations to generate correlated views of the input samples. We compare the performance of the classifier in [15], which analyses three supervised training schemes to train neural networks for HSI multi-label classification. The contributions of our paper are as follows:

1) We introduce a self-supervised contrastive learning approach for feature extraction in hyperspectral remote sensing data that reduces the need for labelled input samples. In contrast to prior works that rely on intricate architectures such as varying depths of ResNet [16], DenseNet [17], AlexNet [18] or Siamese Networks [19] our method engages a straightforward encoder architecture. Despite this streamlined simplicity, the encoder effectively generates rich feature representations that when fine-tuned with a classifier, surpasses existing state-of-the-art methods.

2) While prior studies have conducted multi-label predictions on hyperspectral remote sensing images using patch datasets [15], to the best of our knowledge, the implementation of a self-supervised contrastive learning approach towards this particular task has not been previously addressed.

The remaining of this paper is organised as follows: Section 2 positions our research w.r.t. existing efforts. Section 3 describes the proposed method. In Section 4, we validate our method on HSI patch datasets. Finally, we put forward our concluding remarks and future work prospects in Section 5.

## 2. RELATED WORK

Our work lies at the intersection of two key dimensions.

**Self-supervised contrastive learning (SSCL)**. It is a technique to learn representations of data without labels. The main idea behind SSCL is to draw together representations of augmented views of the same sample while distancing hidden representations derived from different samples [20]. It is applied to various tasks such as image and video classification [21], object detection [22], natural language processing [23], and many other domains. In [14], a simple framework for contrastive learning of visual representations is proposed, *SimCLR*, which is independent of the underlying architecture. Among the significant findings of this work is the pivotal role that strong data augmentation plays in contrastive prediction tasks. Moreover, using a non-linear transformation between the hidden representation and the contrastive loss improves the quality of the learned representations.

Similarly, we position our research in the same territory to learn the relationships hidden within the remote sensing hyperspectral data. We do this by establishing a pretext task of learning a general representation/embedding without engaging annotations. For that, we train an encoder to produce similar representations for the augmented views of the same HSI patch sample.

**Self-supervised Hyperspectral image analysis.** Deep learning methods have demonstrated considerable success in hyperspectral image analysis. However, the lack of sufficient labelled training data remains a significant challenge, as it can result in overfitting and limited generalisation of the models. One approach to overcome this challenge is by constructing deep learning models for hyperspectral image classification that are specifically designed to work with a limited number of labelled samples [24]. While "few-shot" classification can significantly reduce the time and labor required for data collection and labeling, these models are still susceptible to overfitting and limited generalisation. [25] explores the use of supervised contrastive learning (SCL) as a pre-training strategy for HSI classification. In this approach, a feature encoder is pre-trained within a supervised framework using a combination of positive and negative samples, optimising the model parameters in a pairwise manner. However, the scarcity of labelled data remains a persistent challenge.

Alternatively, self-supervised learning in HSI analysis holds promise for surpassing other techniques by enabling representation learning independently of labelled data. [26] introduces a contrastive self-supervised learning (CSSL) algorithm based on Siamese networks [19]. This approach extracts features from pairs of samples and fine tunes a classification model with labelled data for pixel-level, single-label classification. Considering the abundance of unlabelled data, [27] proposes a contrastive learning method for HSI classification. The approach uses a large number of unlabelled samples and employs data augmentation techniques. [28] proposes an unsupervised feature learning method based on autoencoders and contrastive learning. The method aims to extract better features for pixel-level hyperspectral image classification. In [29], hyperspectral change detection is addressed through a self-supervised hyperspectral spatial-spectral understanding network (HyperNet). The latter achieves pixel-wise feature representation without pixel-level annotations. To address the limited availability of labelled pixels in hyperspectral remote sensing images, [30] proposes an architecture that leverages cross-domain convolutional neural networks. This architecture incorporates shared parameters to learn representations across different hyperspectral datasets with varying spectral characteristics and no-pixel level annotations.

Similar to the preceding works, our research focuses on the self-supervised contrastive learning technique to perform hyperspectral image analysis. However different from the above, the downstream task we aim to achieve is that of patch-level, multi-label prediction.

## 3. MODEL DESIGN AND DESCRIPTION

Figure 1 illustrates our two-stage methodology. In the first stage, we employ contrastive learning to train an encoder, ensuring that augmented views of the same input sample are mapped closely in the latent space. This is achieved by mapping each augmented view through the encoder to generate intermediate representations. These are then projected into a vector space using a neural network with two fully-connected layers, where a contrastive loss is applied to fine-tune the similarity. In the second stage, the projection network is removed, and the pre-trained encoder is integrated with a classifier. The entire architecture is then fine-tuned using labelled samples (see Section 4.2). As result, the encoder yields low-dimensional feature

representations, while the classifier discriminates among them.

### 3.1. Model Architecture

**Contrastive Learning Feature Extraction Network.** It comprises several components. The augmentation component applies random vertical and horizontal transformations to an input sample, creating two views known as a positive pair. Each view passes through a network consisting of fully connected layers with Rectified Linear Unit (ReLU) activation and dropout layers. This network component serves as the encoder, responsible for extracting informative features from the two augmented views of the same input sample and mapping them into two hidden (intermediate) representations $h_1$ and $h_2$. The mapping function is defined as $h_i{=}f(W_h \cdot x_i + b_h)$ where $x_i$ represents an augmented view of the input samples $X^M$ ($M =$ the total number of patches), and $W_h$ and $b_h$ being the weights and the bias of the encoder, respectively. The encoder will preserve the spatial dimension of patches yet it will reduce the spectral dimension.

The projection head, is a *neural network* composed of two fully-connected layers with ReLU and dropout layers. Its purpose is to project the hidden representations, $h_1$ and $h_2$ , into another dimensional space represented by $z_1$ and $z_2$. The mapping function is defined as $z_i = g(W_z \cdot h_i + b_z)$ where $z_i$ denotes the vector representation of the intermediate representation, and $W_z$ and $b_z$ represent the weights and the bias of the projection network, respectively. During training, the contrastive loss function will direct the weights to update towards maximising the similarity between the two vector representations $(z)$. This forces the encoder to produce two similar hidden representations $(h)$ for the augmented views generated from the same sample. Ultimately, enabling it to learn relevant features present in the input data. In this context, we employ the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) [31, 32], as our contrastive loss function. The objective of NT-Xent (Eq. 1) is to increase the similarity between the two augmented views (positive pair), while simultaneously asserting their dissimilarity with the negatives samples. The negative samples comprise the remaining augmented views generated from the other input samples within the batch. Following the approach in [14], let $N$ represents the batch size, given a positive pair the remaining $2(N{-}1)$ views serve as negative pairs. Prior to starting the training process for the contrastive learning model, we carefully examined the patches in our dataset removing the duplicates that appeared within each batch.

$$l_{i,j} = -\log \frac{exp(sim(z_i, z_j)/T)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/T)} \tag{1}$$

In Eq. 1, $sim(z_i, z_j)$ is the cosine similarity $\frac{z_i^T z_j}{\|z_i\| \|z_j\|}$, $N$ represents the batch size and since contrastive learning involves two views of each patch, there are $2N$ data points. The function $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ evaluates to 1 if $k$ is not equal to $i$ and 0 otherwise. The temperature scale $T$ is a constant to scale the cosine similarity values ensuring they are not too large nor too small. The similarity between the augmented samples is calculated pairwise for $(i, j)$ projections, encompassing all $2N$ projections within the entire batch. The overall loss function, used for back-propagation is the average taken across all positive pairs in the batch:

$$Loss = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \tag{2}$$

**Downstream network**. After training the base encoder using a contrastive learning approach, we proceed to utilise it for the classification task. At this point, we combine the trained encoder with a classifier. We retain the same architecture of the classifier that was utilised in [15], which is designed to perform patch-level classification on hyperspectral data, with fully-connected layers, non-linear activation function and dropout layers. The classifier takes the hidden representation $h_i$, of the input sample, generated by the base encoder and is trained to predict the class(es) associated with each sample. Towards this end, we employ the Binary cross Entropy with Logits Loss (Eq. 3) as the objective function of the classification model.

$$\ell_c(\hat{y}, y) = L_c = \{l_{1,c}, \ldots, l_{N,c}\}^\top \tag{3}$$

where $n$ indicates the sample index within the batch, and $c$ represents the class label. The individual loss $l_{n,c}$ for each sample $n$ and class $c$ is:

$$\begin{aligned} l_{n,c} = -w_{n,c}[p_c y_{n,c} \cdot \log \sigma(\hat{y}_{n,c}) + (1 - y_{n,c}) \cdot \\ \log(1 - \sigma(\hat{y}_{n,c}))] \end{aligned} \tag{4}$$

In this equation, $p_c$ is the weight of the positive outcome for class $c$, $y_{n,c}$ and $\hat{y}_{n,c}$ denote the ground truth and predicted labels, respectively. $\sigma(\hat{y}n, c)$ represents the sigmoid activation applied to the predicted label and $w_{n,c}$ represents the weight for sample $n$ and class $c$. The label space is defined as $y = \{0, 1\}^c$.

## 4. EXPERIMENTS

In this section, we present the results achieved from our experiments. We utilise two publicly available datasets, Pavia University (PaviaU) and Salinas [33] each of size $X \in \mathbb{R}^{h,w,b}$, with $h, w$, and $b$ representing the height, the width, and the number of spectral bands, respectively. The dimensions of the PaviaU dataset are $610{\times}340{\times}103$ with 9 classes along with a background class. The dimensions of the Salinas dataset are $512{\times}217{\times}204$ with 16 classes along with a background class. From those datasets, we extract patches of size $p \in \mathbb{R}^{h' \times w' \times b}$, thus reducing the spatial dimension while preserving the spectral bands. The resulting patch has a size of $(3, 3, bands)$. We assign multiple labels to these patches to indicate the presence of different classes. The PaviaU multi-label patches dataset exhibits complexity and diversity with $55\%$ of the patches having labels that correspond to mixed classes. In comparison, the Salinas dataset has $21\%$ of patches with mixed class labels.

### 4.1. Implementation Details

The contrastive based-model was trained using unlabelled dataset comprising patches of remote sensing scenes. Geometric transformations including horizontal and vertical flips, generated two views of each input sample. After training the base model, the projection head was removed and a classifier was attached. The new model was trained on the same datasets after associating the hidden representations with the multi-labels. Z-score normalization was applied to the input data. For training the base encoder, the data was split into $90\%$-$10\%$ train and validation sets. For training the classifier, the data was split into train, valid and test sets in the ratio of $\approx 80\%$, $10\%$ and $10\%$ respectively. We used Adam optimisation [34] and employed StepLR learning rate scheduling technique to decay the learning rate by $\gamma{=}0.9$ every 10 epochs. Batch size and epochs were carefully selected to optimise the performance of each baseline.

### 4.2. Performance across different training schemes

In this experiment, we conducted patch-level, multi-label classification of HSI patches. For that purpose we trained the classifier in con-

junction with the CL-based encoder and evaluated the performance under two scenarios: 1) CL-freeze: the layers of the base encoder were frozen and weights were not trainable. 2) CL-tune: The layers of the base encoder were not frozen, allowing the weights to retrain. Performance is reported in terms of average accuracy metric [35].

In the subsequent analysis, we compare the results with those obtained from employing three different schemes, described in [15]. 1) *Iterative* scheme where the autoencoder and classifier function as separate architectures with distinct objectives. It bears resemblance to the iterative training employed in adversarial models. 2) *Joint* scheme [36], the autoencoder and classifier are merged to a unified algorithm. 3) Lastly, in the *Cascade* scheme [37], an autoencoder is trained independently to reconstruct the input and subsequently, the encoder is fine-tuned with a classifier. All schemes exhibit a comparable encoder architecture, featuring a hidden layer of 32 neurons. The classifier architecture is also similar, but the output layer is adjusted to accommodate the varying number of classes in each dataset. No direct comparison was conducted with contrastive learning methods for HSI in the literature. Given their engineered design for pixel-level, single-label classification task, adapting these methods for our patch-level, multi-label classification task would require extensive modifications. In Table 1, our method shows superior performance

**Table 1**. Multi-label Classifier: model accuracy performance

| | *CL-freeze* | *CL-tune* | *Iterative* | *Joint* | *Cascade* |
|---|---|---|---|---|---|
| ***PaviaU*** | 70.56% | **87.87%** | 84.03% | 86.14% | 83.50% |
| ***Salinas*** | 74.90% | **88.86%** | 87.61% | 86.40% | 86.47% |

compared to other schemes when the CL-based encoder is retrained with the classifier (*CL-tune*). Those results are taken from the test set. On PaviaU, the improved performance ranged from 1.73% to 4.37%. For Salinas this improvement ranged from 1.25% to 2.46%. Considering the complexity and diversity of the mixed patches in the PaviaU dataset, it can be concluded that the contrastive learning technique is particularly well-suited for modeling such data, especially given its limited size. Notably, the *CL-tune* variant outperforms the *CL-freeze* variant by a significant margin of 17.31% and 13.98% on the PaviaU and Salinas, respectively. This suggests that during the retraining process, the weights of the encoder update to better align with the features learned by the classifier. These features capture the relevant information the classifier uses for label prediction. Feedback in terms of prediction error is then back-propagated to the earlier layers of the model, namely those of the encoder. This scenario offers the advantage of using the trained weights of the base encoder as a better initialization point for learning/updating the representations of the base encoder.

In addition to outperforming the supervised training schemes, the *CL-tune* scheme exhibits a similar computational profile. As indicated in Table 2, the *CL-tune*-based classifier has fewer trainable parameters compared to the supervised *Joint* training scheme. Consequently, this results in reduced computational requirements.

**Table 2**. Computational requirements

| | ***PaviaU*** | | ***Salinas*** | |
|---|---|---|---|---|
| | ***Joint*** | ***CL-tune*** | ***Joint*** | ***CL-tune*** |
| ***Trainable Params*** ($1\times10^6$) | 6.23 | 6.21 | 6.25 | 6.22 |
| ***forward/backwardpass (MB)*** | 14.65 | 14.12 | 10.48 | 8.92 |
| ***Params(MB)*** | 24.92 | 24.85 | 25 | 24.89 |
| ***Estimated Total size(MB)*** | 40.31 | 39.93 | 36.43 | 35.01 |

### 4.3. Impact of the dimension of the hidden representation

In Section 4.2 we utilised a contrastive learning base encoder to produce hidden representations of dimension 32. However, considering the patches datasets we are using, the encoder compressed the spectral depth from $(3\times3\times103)$ to $(3\times3\times32)$ for PaviaU and from $(3\times3\times204)$ to $(3\times3\times32)$ for Salinas. To investigate the potential consequences of this spectral compression, we conducted an experiment where we increased the hidden representation dimension to 64. We then retrained our contrastive learning base encoder on patches of both datasets. The pre-trained base encoder and the classifier were subsequently fine-tuned on the labelled data employing either the freezing or retraining of the encoder's layers.

Table 3 presents the impact of the hidden representation dimension on performance. Increasing the size to 64 neurons resulted in improvements in both the *CL-tune* and the *CL-freeze* schemes. Specifically, the the *CL-freeze* showed improvement of $\approx4\%$ and $2.3\%$ for the PaviaU and the Salinas datasets, respectively. The self-supervised contrastive learning approach allowed the encoder to better capture similarities in the data by preserving more information in the hidden representation. This reaffirms that contrastive learning methods effectively overcome limitations posed by the size and complexity of the available data, as evidenced by the results obtained from the PaviaU patches dataset.

**Table 3**. Multi-label Classifier: model accuracy performance w.r.t higher dimensional hidden representation

| | ***CL-freeze*** | ***CL-tune*** | ***CL-freeze*** | ***CL-tune*** |
|---|---|---|---|---|
| | ***hidden layer 32*** | | ***hidden layer 64*** | |
| ***PaviaU*** | 70.56% | **87.87**% | 74.06% | **88.45**% |
| ***Salinas*** | 74.90% | **88.86**% | 77.13% | **89.74**% |

## 5. CONCLUSION

This paper presents a two-stage method for training a multi-label classifier for hyperspectral remote sensing images. The method, employs a minimalist architecture bolstered by self-supervised contrastive learning to pre-train a base encoder, thereby generating meaningful hidden representations that serve as input for the classifier. Results indicate that our approach outperforms fully supervised training methods in the task of multi-label classification on hyperspectral remote sensing images. Utilising both labelled and unlabelled data, we fine-tune the classifier while capitalising on the base encoder's unsupervised training. This contrasts with end-to-end supervised learning approaches which often suffer from overfitting and yield minimal generalisation due to their reliance on limited labelled datasets. Future research will aim to broaden the empirical validation across diverse hyperspectral datasets, and explore model interpretability for enhanced and more transparent decision-making. Additionally, tackling spectral variability challenges in hyperspectral unmixing through contrastive learning constitutes another promising research avenue.

**Acknowledgments**: This work is supported by Flanders Innovation & Entrepreneurship-VLAIO, under grant no. HBC.2020.2266.

## 6. REFERENCES

[1] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang, "Self-supervised learning: Genera-

tive or contrastive," *TKDE*, 2023.

[2] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *TPAMI*, 2020.

[3] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *ICML*, 2019.

[4] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon, "A survey on contrastive self-supervised learning," *Technologies*, 2020.

[5] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton, "Contrastive representation learning: A framework and review," *Access*, 2020.

[6] Kriti Ohri and Mukesh Kumar, "Review on self-supervised image recognition using deep neural networks," *Knowledge-Based Systems*, 2021.

[7] Ting Chen, Calvin Luo, and Lala Li, "Intriguing properties of contrastive losses," *NIPS*, 2021.

[8] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *TGRS*, 2020.

[9] Zhonghao Chen, Guoyong Wu, Hongmin Gao, Yao Ding, Danfeng Hong, and Bing Zhang, "Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation," *Expert Systems with Applications*, 2023.

[10] Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework," *TGRS*, 2023.

[11] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu, "Deep recurrent neural networks for hyperspectral image classification," *TGRS*, 2017.

[12] Xiaofei Yang, Yunming Ye, Xutao Li, Raymond YK Lau, Xiaofeng Zhang, and Xiaohui Huang, "Hyperspectral image classification with deep learning models," *TGRS*, 2018.

[13] Onuwa Okwuashi and Christopher E Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognition*, 2020.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[15] Salma Haidar and José Oramas, "Training methods of multi-label prediction classifiers for hyperspectral remote sensing images," *arXiv preprint arXiv:2301.06874*, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, 2017.

[19] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al., "Siamese neural networks for one-shot image recognition," in *ICML*. Lille, 2015.

[20] Zixin Wen and Yuanzhi Li, "Toward understanding the feature learning process of self-supervised contrastive learning," in *LCML*, 2021.

[21] Haiping Wu and Xiaolong Wang, "Contrastive learning of image representations with cross-video cycle-consistency," in *ICCV*, 2021.

[22] Wei Wu, Hao Chang, Yonghua Zheng, Zhu Li, Zhiwen Chen, and Ziheng Zhang, "Contrastive learning-based robust object detection under smoky conditions," in *CCVPR*, 2022.

[23] Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau, "Contrastive data and learning for natural language processing," in *NAC-ACL*, 2022.

[24] Sen Jia, Shuguo Jiang, Zhijie Lin, Nanying Li, Meng Xu, and Shiqi Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, 2021.

[25] Lingbo Huang, Yushi Chen, Xin He, and Pedram Ghamisi, "Supervised contrastive learning-based classification for hyperspectral image," *Remote Sensing*, 2022.

[26] Lin Zhao, Wenqiang Luo, Qiming Liao, Siyuan Chen, and Jianhui Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE-GRSL*, 2022.

[27] Sikang Hou, Hongye Shi, Xianghai Cao, Xiaohua Zhang, and Licheng Jiao, "Hyperspectral imagery classification based on contrastive learning," *TGRS*, vol. 60, 2021.

[28] Zeyu Cao, Xiaorun Li, Yueming Feng, Shuhan Chen, Chaoqun Xia, and Liaoying Zhao, "Contrastnet: Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral imagery classification," *Neurocomputing*, 2021.

[29] Meiqi Hu, Chen Wu, and Liangpei Zhang, "Hypernet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *TGRS*, 2022.

[30] Hyungtae Lee and Heesung Kwon, "Self-supervised contrastive learning for cross-domain hyperspectral image representation," in *ICASSP*, 2022.

[31] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.

[32] Wilhelm Ågren, "The nt-xent loss upper bound," 2022.

[33] `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`, "Hyperspectral remote sensing scenes," *Computational Intelligence Group CIC*, 2011.

[34] Diederik Kingma and Jimmy Ba., "Adam: A method for stochastic optimization," *"ICLR"*, 2014.

[35] Mohammad S Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University*, 2010.

[36] Yisen Liu, Songbin Zhou, Hongmin Wu, Wei Han, Chang Li, and Hong Chen, "Joint optimization of autoencoder and self-supervised classifier: Anomaly detection of strawberries using hyperspectral imaging," *CEA*, 2022.

[37] Chen Xing, Li Ma, Xiaoquan Yang, et al., "Stacked denoise autoencoder based feature extraction and classification for hyperspectral images," *Journal of Sensors*, 2016.