Enhancing Contrastive Learning With Positive Pair Mining for Few-Shot Hyperspectral Image Classification

Nassim Ait Ali Braham, Julien Mairal, Jocelyn Chanussot, Lichao Mou, Xiao Xiang Zhu

Enhancing Contrastive Learning With Positive Pair Mining for Few-Shot Hyperspectral Image Classification

Nassim Ait Ali Braham, Julien Mairal, Jocelyn Chanussot, *Fellow, IEEE*, Lichao Mou, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—In recent years, deep learning has emerged as the dominant approach for hyperspectral image (HSI) classification. However, deep neural networks require large annotated datasets to generalize well. This limits the applicability of deep learning for real-world HSI classification problems, as manual labeling of thousands of pixels per scene is costly and time consuming. In this article, we tackle the problem of few-shot HSI classification by leveraging state-of-the-art self-supervised contrastive learning with an improved view-generation approach. Traditionally, contrastive learning algorithms heavily rely on hand-crafted data augmentations tailored for natural imagery to generate positive pairs. However, these augmentations are not directly applicable to HSIs, limiting the potential of self-supervised learning in the hyperspectral domain. To overcome this limitation, we introduce two positive pair-mining strategies for contrastive learning on HSIs. The proposed strategies mitigate the need for high-quality data augmentations, providing an effective solution for few-shot HSI classification. Through extensive experiments, we show that the proposed approach improves accuracy and label efficiency on four popular HSI classification benchmarks. Furthermore, we conduct a thorough analysis of the impact of data augmentation in contrastive learning, highlighting the advantage of our positive pair-mining approach.**

***Index Terms*—Contrastive learning, hyperspectral image (HSI) classification, positive pair mining, self-supervised learning.**

I. Introduction

HYPERSPECTRAL imagery is a very important and powerful technology in remote sensing. Hyperspectral images (HSI) contain hundreds of narrow bands covering a wide range of the electromagnetic spectrum, from visible to near infrared, thereby capturing very rich information about the physical characteristics of objects in a scene. This unique property of hyperspectral data, coupled with an increasing availability of cost-effective sensors, and an improving spatial and spectral resolution, has enabled many applications in agriculture, environmental monitoring, and biomedical imaging, to name a few. This rapid advancement in measurement instruments has significantly increased the amounts of data available, subsequently calling for the development of HSI analysis algorithms, especially for HSI classification.

The problem of HSI classification consists in assigning a semantic label to every pixel in an HSI. It is an important and challenging problem that has received considerable attention from the remote sensing community in the last decades. Progress in this field has been closely intertwined with advancements in machine learning. In early works, traditional HSI classification methods followed on a two-stages procedure. First, they employed a hand-crafted feature extraction and/or dimensionality reduction algorithm such as local binary patterns [1], histogram of oriented gradients [2], principal component analysis (PCA) [3], or independent component analysis [4]. Then, a traditional classification algorithm such as support vector machines (SVMs) [5], multinomial logistic regression [6], or random forests [7] was learned on the manually generated features. Nowadays, with the rise of deep learning, neural networks have emerged as the dominant approach. Their capability to learn latent representations from raw data in an end-to-end manner eliminates the need for extensive feature engineering.

Deep learning has enabled tremendous progress in HSI analysis in the last few years, especially for HSI classification [8]. For example, Hu et al. [9] used a 1-D convolutional neural network for spectral feature extraction. Zhong et al. [10] proposed a spectral–spatial residual network (SSRN) leveraging 3-D convolutions for enhanced feature extraction. Hamida et al. [11] introduced a 3-D architecture for HSI classification. Mou et al. [12] designed a recurrent neural network for HSI classification considering the spectrum as a sequence. Hong et al. [13] proposed a novel spectral–spatial transformer backbone by grouping adjacent spectral bands and generating tokens in the spectral dimension. Hong et al. [14] adapted graph

Manuscript received 30 October 2023; revised 31 January 2024; accepted 13 February 2024. Date of publication 4 March 2024; date of current version 22 April 2024. This work was supported in part by the DLR-DAAD Research Fellowships programme (Scholarship ID: 57540125), and in part by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. An earlier version of this article was presented at the International Geoscience and Remote Sensing Symposium (IGARSS) 2022 [DOI: 10.1109/IGARSS46834.2022.9884494]. *(Corresponding author: Nassim Ait Ali Braham.)*

Nassim Ait Ali Braham is with the Chair of Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany, and also with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: Nassim.AitAliBraham@dlr.de).

Lichao Mou and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany.

Julien Mairal and Jocelyn Chanussot are with the Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Digital Object Identifier 10.1109/JSTARS.2024.3371909

convolutional networks (GCNs) for HSI classification and proposed miniGCN, an enhanced GCN model that supports mini-batch training, enabling training on larger images. However, despite these efforts, training deep neural networks following the supervised learning paradigm requires large and well-annotated datasets to generalize and avoid overfitting. Unfortunately, accurately labeling thousands of pixels in an HSI is time consuming, and requires expert knowledge. This limitation severely impedes the applicability of deep learning on real-world hyperspectral datasets. Therefore, label-efficient learning is a critical research direction that requires further investigation to enable the use of deep learning in HSI analysis.

Significant efforts have been devoted to develop label efficient learning methodologies, such as semisupervised learning [15], meta-learning [16], weakly supervised learning [17], and unsupervised/self-supervised learning [18]. These paradigms have been exploited in remote sensing as well to reduce the need for manual annotation. In particular, several HSI classification works have shown promising results with sparse training sets [19]. Yet, the problem of HSI classification with limited labels remains open. Semi-supervised learning approaches make use of unlabeled samples in conjunction with annotated pixels, often relying on consistency losses or pseudolabeling techniques [20]. On the other hand, few-shot meta-learning methods pretrain classification models on large labeled datasets in a way that enables the model to quickly adapt to new previously unseen classes [21]. Earlier works in unsupervised learning rely on classical representation learning algorithms such as autoencoders [22] to mitigate overfitting.

Of particular interest to us, self-supervised learning algorithms have led to significant breakthroughs in computer vision in the last few years [23]. These methods aim at learning versatile representations from unlabeled data [24] and leverage the resulting pretrained models for various downstream tasks. It has been shown that such pretraining strategies can even outperform supervised pretraining on several problems [25]. One specific family of methods that has driven progress in this field is contrastive learning.[1] Contrastive methods train a model to produce similar representations for two randomly augmented versions of the same image. Enforcing such invariance results in semantically rich representations provided that the augmentations are well-chosen.

In this article, we leverage recent advancements in self-supervised contrastive learning to enable accurate HSI classification with limited labels. Specifically, we pretrain a convolutional residual backbone using a state-of-the-art self-supervised learning algorithm, namely Barlow-Twins (BT) [26], on unlabeled pixels in a HSI. The resulting pretrained model can be efficiently adapted following a lightweight classification protocol with limited labels. In addition, we propose to leverage interdependencies between pixels in the scene to design positive pair-mining strategies, providing high-quality views for contrastive learning algorithms. Our experiments demonstrate that incorporating diverse samples as positive pairs enriches the supervision signal and enhances robustness to data augmentation. Finally, we conduct extensive ablation studies to analyze the impact of view generation, including the choice of data augmentation, on the performance of contrastive learning for HSI classification. Our approach is compatible with any contrastive learning method and can be applied to both spectral and spatial–spectral classifiers. The results demonstrate that the proposed pipeline is superior to plain supervised learning, few-shot learning algorithms and classical contrastive pretraining. Our contributions can be summarized as follows.

1) We propose a self-supervised learning approach for HSI classification with limited labels by leveraging state-of-the-art contrastive learning algorithms.
2) We propose two positive pair-mining strategies for contrastive learning to exploit the interdependencies between different pixels/patches in a scene and reduce the dependence on high-quality data augmentations in the hyperspectral domain.
3) We evaluate the proposed approach on four popular HSI classification datasets and analyze the impact of the view generation strategy, pair mining, and data augmentation, on the quality of the representations learnt.

The rest of this article is organized as follows. In Section II, we discuss the background and related work from the HSI classification and self-supervised learning literature. In Section III, we present the proposed methodology. In Section IV, we describe our experimental setting and results. Finally, Section V concludes this article.

II. Related Work

A. HSI Classification With Limited Labels

Label efficient HSI classification has been a very active research topic in the past few years [19]. Numerous studies have employed various learning paradigms and tailored network architectures toward a common objective: producing accurate classification maps with a handful of labels. One line of work has exploited few-shot learning algorithms, e.g., Tang et al. [27] use prototypical networks [28] to train a model on a combination of labeled source HSI classification datasets, enabling its use on target datasets with novel classes in a few-shot setting. Similarly, Gao et al. [29] adapt relation networks [30] and use a learnable distance metric to compare the patches in an HSI. An important drawback of these methods is that they require a large amount of annotated pretraining data and special care when the source and target sensors do not match.

Another successful line of work leverages semi-supervised learning to make use of unlabeled pixels in an HSI and increase label efficiency. For example, Wu et al. [20] employed pseudolabels generated by a Dirichlet process mixture model to train a deep recurrent network for HSI classification with few labels. Wu et al. [31] proposed a semi-supervised approach that combines self-training with a spatial constraint to improve the consistency of pseudolabels.

More recently, researchers have investigated self-supervised learning methodologies for HSI classification. Yang et al. [32]

[1] In this work, we denote by contrastive learning all methods using a joint-embedding/siamese architecture, including algorithms that do not explicitly use negative pairs.

designed a pretext task in which the model is pretrained to predict the scale and flipping labels for a patch. Liu et al. [33] adapted the idea of contrastive multiview coding [34] to HSIs by dividing the spectral channels into two different subsets to create distinct views. Zhao et al. [35] used a contrastive learning method, namely SimCLR [36], to perform HSI classification with a limited number of samples per class.

B. Contrastive Learning

Contrastive learning is a family of self-supervised learning algorithms that train a model to be invariant to input transformations. Specifically, the model learns to produce similar representations for two randomly augmented versions of the same image. The underlying assumption is that augmentations preserve the semantics and only alter the style of the image. However, without a constraint on the diversity of the representations, this objective can result in a trivial solution, known as collapse, where all inputs yield a constant representation. To address this issue, various types of methods have been proposed in the literature.

1) Contrastive methods with negative samples such as SimCLR [36] and MoCo [37] explicitly encourage the representations of different images within the same batch to diverge.
2) Distillation-based methods such as SimSiam [38] or BYOL [39] follow a student–teacher model and introduce asymmetry in the architecture (e.g., stop-gradient and predictor subnetwork).
3) Clustering-based methods such as SwAV [25] introduce equipartition constraints in the embedding space.
4) Regularization-based methods, e.g., Barlow-Twins (BT) [26], explicitly avoid collapse using a regularization term in the loss function that enforces feature decorrelation.

Contrastive learning has gained popularity in HSI analysis. Hou et al. [40] used SimCLR to pretrain and finetune a convolutional neural network on a hyperspectral scene. Hu et al. [41] leveraged the BYOL algorithm with a transformer architecture combined with hand-crafted augmentations for HSI classification. Huang et al. [42] proposed a multiscale contrastive learning approach with a 3-D Swin transformer backbone. Zhu et al. [43] proposed a multiscale approach based on SimCLR with a novel light-weight backbone for HSI classification. Li et al. [44] leveraged BYOL with an occlusion augmentation to learn robust representations. Xue et al. [45] exploit contrastive learning with a variable number of views in a multitask setting for multimodal few-shot land cover classification. Guan et al. [46] use contrastive learning for HSI classification with two distinct encoder branches: a spatial encoder that takes as input a masked patch, and a spectral encoder that takes as input a partially masked pixel.

C. Positive and Negative Pair Mining in Contrastive Learning

Early works in supervised and unsupervised contrastive learning [47], [48] generally relied on a small number of negative samples per anchor (one or two). Consequently, the quality of the learned representations heavily depended on the selection strategy of the negative samples. Easy negatives were uninformative, while hard samples were challenging to learn. Nowadays, most self-supervised contrastive learning methods use a very large number of negatives (typically all samples within the batch), without any explicit mining strategy. Further analyses have shown that in a large batch, only few negative samples were contributing to the loss during pretraining [49], [50]. Based on that, careful negative pair selection and generation strategies have been proposed for contrastive methods. Nevertheless, negative pair mining is not widely used in modern contrastive learning methods. Positive pair mining has attracted less attention in the unsupervised setting given the difficulty of accurately selecting true positives without labels. Nevertheless, some efforts have been made in that direction. Jean et al. [51] proposed to exploit spatial information to select positive and negative samples with a triplet loss on NAIP and Landsat imagery. The intuition is that nearby patches share common semantics, whereas distant patches should have dissimilar representations. Kang et al. [52] incorporated a similar idea in MoCo to generate more informative positive pairs. SeCo [53] and SSL4EO-S12 [54] used temporal positives, i.e., images of the same location from different seasons captured by the Copernicus Sentinel-2 mission. Dwibedi et al. [55] proposed to select positive pairs using KNN retrieval in the embedding space of the model during pretraining. Wang et al. [56] adopted a similar approach in a multimodal hyperspectral—LiDAR setting.

The effectiveness of contrastive learning relies heavily on the quality of augmentations used for view generation. While these augmentations have been fine tuned for natural imagery, their direct application to hyperspectral imaging poses challenges due to the significant domain shift. The unique spectral characteristics of hyperspectral data call for specialized view generation approaches.

III. Methodology

In this section, we introduce the different components of the proposed method, namely the pretraining algorithm, the classification protocol, and the view generation procedure.

A. Overview

Let $X \in \mathbb{R}^{N\times M\times C}$ be an HSI. We denote by $x \in \mathbb{R}^{p\times p\times d}$ a hypercube of size $p \times p$ sampled from the scene. Our approach follows a two-stage process involving pretraining and classification. Initially, we employ a self-supervised contrastive learning algorithm to pretrain an encoder f on all pixels in the scene, including unlabeled ones. Then, we utilize the encoder f to train a classification model g using a limited set of labeled pixels. During classification, we predict the label of the central pixel in the patch x. This approach effectively mitigates overfitting while leveraging the knowledge gained from the vast number of unlabeled pixels. The overall method is depicted in Fig. 1.

B. Pretraining Algorithm

During the pretraining stage, we treat all pixels in the scene as unlabeled samples and train a convolutional encoder f using a state-of-the-art self-supervised learning algorithm, namely BT [26]. Similar to other contrastive methods, the pretraining

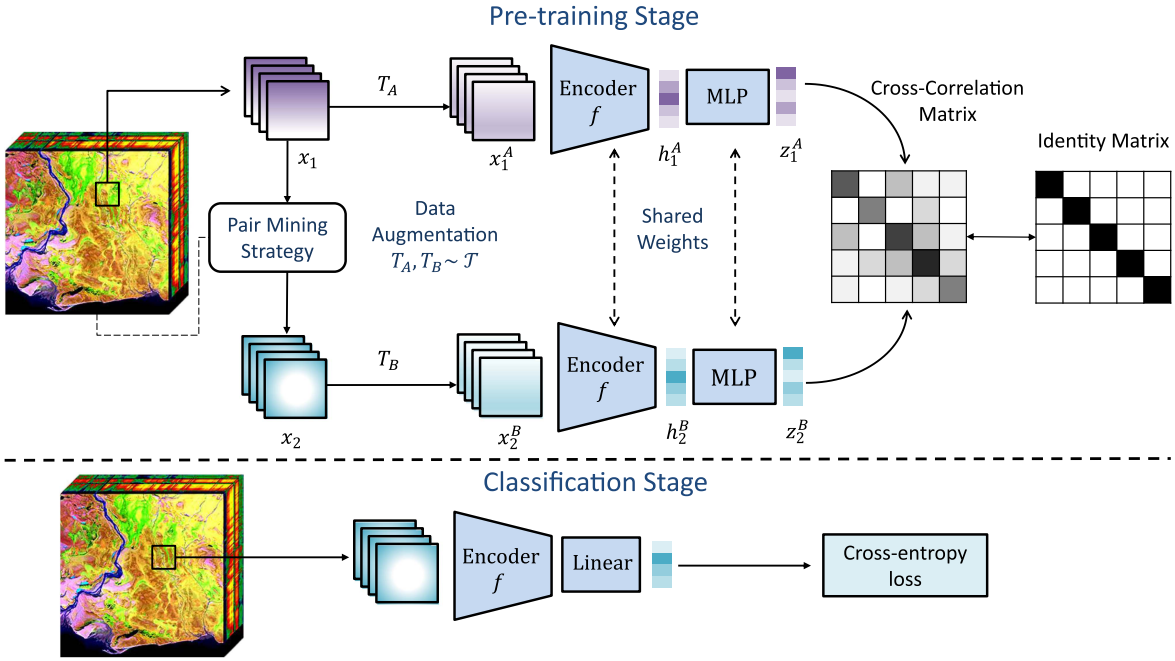


Fig. 1. Overview of the proposed self-supervised learning approach for HSI classification.

objective aims at learning similar embeddings for distorted versions of the same input. In addition, BT prevents collapse by minimizing the redundancy between the features in the embedding space. Specifically, from an input x, two views $x^A = T_A(x)$ and $x^B = T_B(x)$ are stochastically generated using two sets of augmentations T_A and T_B. In this work, we also leverage the global context in the scene to generate positive pairs using different patches x_1 and x_2 (see Section III-C). For the sake of simplicity, we denote the two views by x_A and x_B here. The views are fed into the shared encoder f to obtain their representations $h^A = f(x^A)$ and $h^B = f(x^B)$. Subsequently, an additional multilayer perceptron (MLP) head is applied to further project the representations into the embedding space, resulting in the embedding vectors $z^A = \text{MLP}(h^A)$ and $z^B = \text{MLP}(h^B)$. Since x^A and x^B form a positive pair, their embeddings should be similar. Furthermore, following the redundancy minimization principle, the dimensions of z^A and z^B should be decorrelated. Let Z^A and Z^B be the embedding matrices for a batch of positive pairs. To quantify the degree of correlation, we compute the cross-correlation matrix C along the batch, defined as

$$\mathcal{C}_{ij} = \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}. \tag{1}$$

The objective is to make C close to identity. This is achieved by optimizing the following loss function:

$$\mathcal{L} = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance to augmentation}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction}} \tag{2}$$

Algorithm 1: Pseudo-Code for BT with Positive Pair Mining.

```
Initialize Backbone f, Projection Head MLP
Define a selection strategy for positive pair mining (e.g.,
 neighboring patches and superpixels-based)
Define augmentations T_A and T_B
  for each batch do
    for i = 1, ..., N do
      Sample a patch x_i from the HSI
      Sample a positive pair (x_i, x'_i) based on the selection
       strategy
      Data augmentations: x_i^A = T_A(x_i), x_i^B = T_B(x'_i)
      Compute representations: h_i^A = f(x_i^A), h_i^B = f(x_i^B)
      Compute projected embeddings:
       z_i^A = MLP(h_i^A), z_i^B = MLP(h_i^B)
    end for
    Compute cross-correlation matrix C across the batch
    Compute the Loss
     L = sum_i (1 - C_ii)^2 + λ sum_i sum_{j≠i} C_ij^2
    Take a gradient step to minimize L
  end for
```

where $\lambda \in \mathbb{R}^+$ is a tradeoff parameter. A pseudocode summarizing the pretraining stage is provided in Algorithm 1.

C. Classification Stage

The pretraining step yields an encoder f that learns informative representations for the pixels in the scene. Note that the MLP projector is no longer used after pretraining. In the second step, our goal is to leverage this pretrained model to construct a

classifier g using a limited number of labeled samples. For this purpose, we explore several approaches.

1) *Linear classification:* This approach involves constructing a linear classifier l on top of the representations produced by the frozen encoder f. The resulting classifier is denoted by $g = l \circ f$. Due to its simplicity, computational efficiency, and effectiveness, we conduct most of our experiments with the linear classification protocol.
2) *Support vector machines (SVM):* This approach is similar to the linear protocol, but instead of appending a linear layer to the frozen backbone, we train an SVM classifier [5].
3) *Multilayer perceptron (MLP):* In this protocol, we introduce additional capacity in the classification stage by appending an MLP network to the frozen backbone. This allows the model to learn more complex decision boundaries when the representations from the pretraining stage are not linearly separable.
4) *Finetuning:* The finetuning protocol utilizes the encoder's weights as an initialization to train the classifier $g = l \circ f$. All the parameters of f are trainable. Therefore, special care must be taken to avoid overfitting or deviating too far from the initial pretrained model. To mitigate these risks, we first train the linear layer l following the linear protocol before fine tuning the backbone. This helps prevent large gradients that may result from the random initialization of l.

The choice of the classification protocol heavily depends on the amount of labels available in the downstream classification phase. Given our focus on a low-shot setting in this work, we prioritize lightweight classifiers to mitigate overfitting.

D. View Generation

One of the most critical components in contrastive learning is data augmentation. Prior research has shown that many state-of-the-art self-supervised learning algorithms perform similarly when evaluated in a fair setting with the same hyperparameters, such as momentum encoder, projector size, and optimizer [57]. However, the choice of augmentations used to generate the positive pairs can dramatically impact the quality of the representations, as demonstrated in [36]. Two essential classes of augmentations in contrastive learning are random cropping and color distortions (such as color jittering). Removing either of these two augmentations can severely hurt downstream performance. In the context of HSI classification, random cropping is limited due to the small spatial context in the patches (or worst case no spatial context with a purely spectral classifier). Moreover, performing color jittering with more than three channels is challenging. Furthermore, spectral information holds greater importance in HSIs compared to natural imagery. This triggers the need for an adapted view generation procedure in HSIs. Therefore, we put our focus on two key aspects: the selection strategy for positive pairs and the choice of data augmentation.

1) Pair-Mining Strategy: Traditionally, contrastive learning methods generate positive pairs by stochastically augmenting the same input x twice. However, since the patches come from a single scene, the samples are not independent from each other. Therefore, we can exploit the relationship between pixels to select positive pairs from the image. This data-driven augmentation strategy reduces the dependence on carefully tuned transforms. In this work, we consider the following two simple yet effective pair-mining strategies.

1) *Neighboring patches:* By incorporating spatial locality, we can sample spatially neighboring patches in the scene, thus leveraging the spatial regularity prior for pretraining. Specifically, instead of augmenting the same patch x, we use x and another nearby patch x' as a positive pair. If the neighborhood is small (e.g., by imposing that x and x' overlap), we can assume that they should have similar representations.
2) *Superpixel guided:* Spatial locality allows to enrich the supervision signal. However, a data-agnostic selection-strategy-based solely on spatial proximity fails to consider larger neighborhoods without introducing many false positives, thereby harming the pretrained model. A simple workaround is to take into account the similarities between the pixels in a scene and cluster them in an accurate manner. One can achieve this with little to no computational overhead by simply segmenting the image into coarse superpixels before pretraining using classical, proven algorithms such as Felzenszwalb's efficient graph-based segmentation algorithm [58] or the simple linear iterative clustering [59] algorithm. Assuming that pixels lying within the same superpixels are similar, one can generate positive pairs by randomly selecting x and x' from the same superpixel.

2) Data Augmentation: In this work, we consider several common augmentations for contrastive learning. These can be grouped into two classes: spatial and spectral transform. A brief description of each augmentation is provided as follows.

1) *Spatial transforms:* This type of augmentations require a sufficiently large spatial context. They preserve spectral information, but they are not applicable to purely spectral classifiers. We experiment with the transformations listed as follows.
   a) *Random flipping:* Perform horizontal or vertical flipping with probability $p = 0.5$.
   b) *Random rotation:* Rotate the patch by 90°, 180°, or 270°.
   c) *Random resized crop:* Select a small crop from the patch and resize it. We use the standard transform implemented in PyTorch.
2) *Spectral Transforms:* Spectral transforms can be used on individual pixels, but may distort the semantics if they are too extreme. In the following, we list the augmentations we experimented within this work.
   a) *Scalar multiplication:* Multiply all the bands of a patch by a scalar value α drawn from a uniform distribution in a range $[\alpha_{\min}, \alpha_{\max}]$.
   b) *Gaussian noise:* Add a Gaussian Noise centered around 0 with standard deviation σ, independently sampled for every channel in the patch.

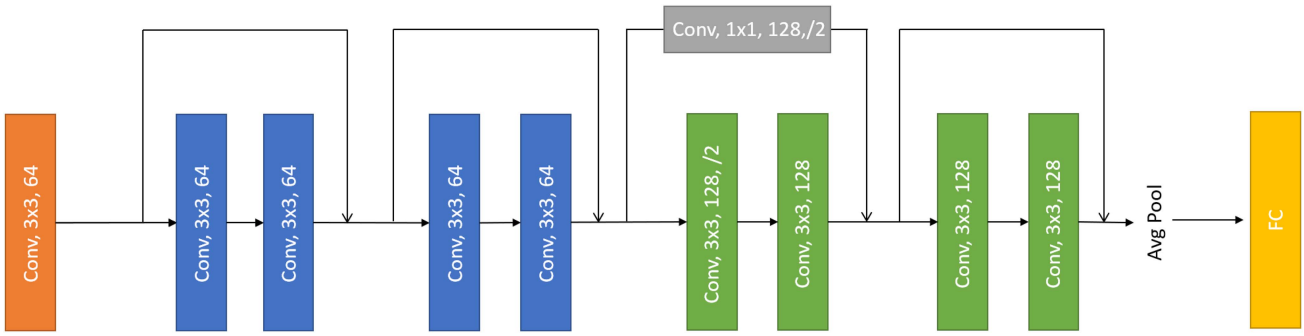


Fig. 2. Architecture of the Mini-Resnet used in this work. It consists of multiple (Conv + Batch Norm + ReLU) blocks with residual connection. Batch Norm and ReLu activation are not displayed for simplicity.

TABLE I
NUMBER OF LABELED PIXELS PER CLASS IN THE PAVIAU DATASET

No	Class Name	Number of Labeled Pixels
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

c) *Random band mask:* Set to 0 a randomly selected number $n \in [n_{\min}, n_{\max}]$ of bands. All spatial locations of the patch are masked in the selected bands.
d) *Random pixel mask:* Set to 0 a randomly selected number $n \in [n_{\min}, n_{\max}]$ of pixels in the patch. All bands are masked for the selected pixels.
e) *Random band swap:* Permute the values of randomly chosen number $n \in [n_{\min}, n_{\max}]$ pairs of adjacent bands.
f) *Random offset:* Add a randomly drawn offset value $b \in [b_{\min}, b_{\max}]$ to all bands in the patch.

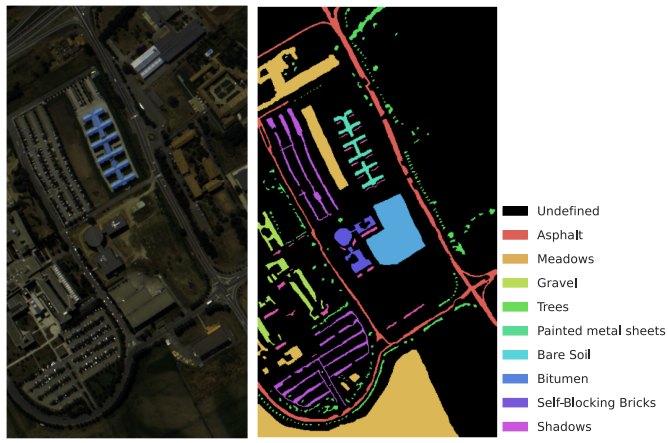
IV. EXPERIMENTS

A. Experimental Setting

In this section, we describe our evaluation setting, including the datasets we use, the methods we evaluate, and the implementation details.

1) HSI Datasets: We evaluate our proposed approach on four widely used HSI classification datasets, namely Pavia University (PaviaU), Salinas, Wuhan HanChuan Hyperspectral Image (WHU-Hi-HanChuan), and the Houston University 2013 dataset. The individual characteristics of each dataset are described as follows.

1) *Pavia University (PaviaU):* The PaviaU dataset was acquired by the ROSIS-03 sensor over an urban area in Pavia, northern Italy. The size of the scene is 610 × 340 pixels with a spatial resolution of 1.3 m. We utilize 103 out of the original 115 spectral bands, ranging from 0.43 to 0.86 μm, after excluding noisy bands. The dataset contains nine classes, with a total of 42 776 labeled pixels in the ground truth. Details about the class distribution are provided in Table I. The RGB visualization of the scene and its corresponding ground truth mask are depicted in Fig. 3.



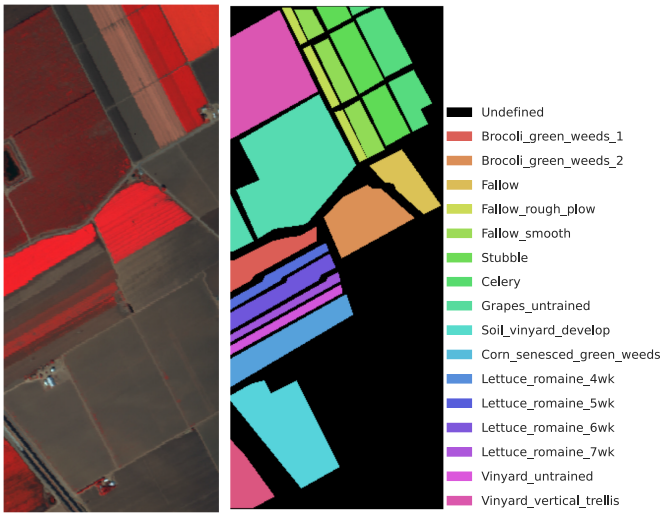
(a) (b)

Fig. 3. PaviaU dataset. (a) RGB. (b) Ground truth.

2) *Salinas:* The Salinas dataset was collected using the AVIRIS sensor, covering the Salinas Valley in California. It comprises a scene of 512 × 217 pixels with a spatial resolution of 3.7 m. The dataset initially includes 224 spectral bands ranging from 0.4 to 2.5 μm. The 20 water absorption bands were removed, leaving a total of 204 bands. The ground truth consists of 16 crop type categories, with a total of 54 129 labeled pixels. The distribution of samples for each class can be found in Table II. RGB visualization of the image along with the ground truth are provided in Fig. 4.
3) *Wuhan HanChuan Hyperspectral Image (WHU-Hi-HanChuan):* The WHU-Hi-HanChuan dataset covers a farming area in Hanchuan, Hubei, China. The image was collected via the Headwall Nano-Hyperspectral imaging sensor mounted on an UAV platform. The scene size is 1217 × 303 pixels, with a spatial resolution of 0.109 m. It comprises 274 spectral bands ranging from 0.4 to 1.0

TABLE II
NUMBER OF LABELED PIXELS PER CLASS IN THE SALINAS DATASET

No	Class Name	Number of Labeled Pixels
1	Brocoli green weeds 1	2009
2	Brocoli green weeds 2	3726
3	Fallow	1976
4	Fallow rough plow	1394
5	Fallow smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes untrained	11271
9	Soil vinyard develop	6203
10	Corn green weeds	3278
11	Lettuce romaine 4wk	1068
12	Lettuce romaine 5wk	1927
13	Lettuce romaine 6wk	916
14	Lettuce romaine 7wk	1070
15	Vinyard untrained	7268
16	Vinyard vertical trellis	1807



(a) (b)

Fig. 4. Salinas dataset. (a) RGB. (b) Ground truth.

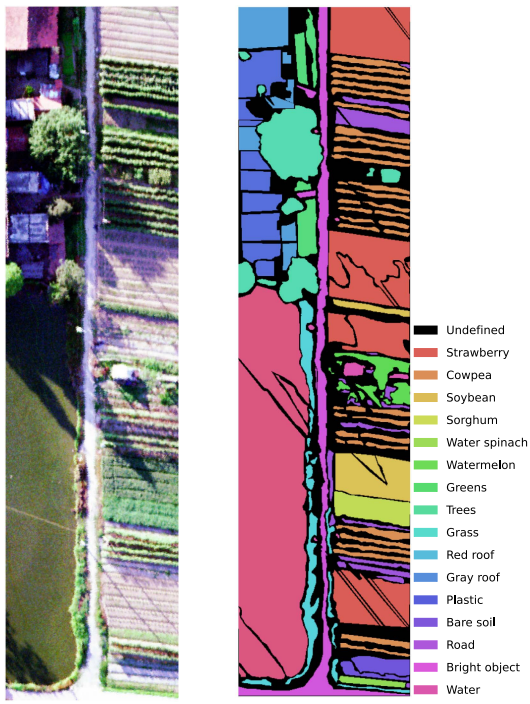
μm. The ground truth consists of 253 580 labeled pixels from 16 different crop type categories. The detailed class distribution is provided in Table III. The RGB image and corresponding ground truth are visualized in Fig. 5.

4) *Houston University 2013:* The Houston dataset was acquired by the ITRES-CASI-1500 sensor over the University of Houston campus and its neighboring urban area. The dataset comprises a scene of 349 $\times$ 1905 pixels with a spatial resolution of 2.5 m. It includes 144 spectral bands, covering a wavelength range from 0.38 to 1.05 μm. The ground truth contains 15 classes of urban land-cover, with a total of 15 104 labeled pixels. The class distribution and sample images are detailed in Table IV and Fig. 6, respectively.

2) Methods: In order to show the effectiveness of pretraining and the importance of a good pair-mining strategy, we compare the proposed approach to the following baselines.

TABLE III
NUMBER OF LABELED PIXELS PER CLASS IN THE WHU-HI-HANCHUAN DATASET

No	Class Name	Number of Labeled Pixels
1	Strawberry	44740
2	Cowpea	22758
3	Soybean	10292
4	Sorghum	5358
5	Water spinach	1205
6	Watermelon	4538
7	Greens	5908
8	Trees	17983
9	Grass	9474
10	Red roof	10521
11	Gray roof	16916
12	Plastic	3684
13	Bare soil	9121
14	Road	18565
15	Bright object	1141
16	Water	75406



(a) (b)

Fig. 5. WHU-Hi-HanChuan dataset. (a) RGB. (b) Ground truth.

1) *Support vector machine (SVM):* A common pixel-level baseline that uses as input the spectral signature of a pixel and trains an SVM) classifier [60].

2) *Extended morphological profile-support vector machine (EMP-SVM):* EMP-SVM [61] is a popular traditional feature extraction baseline. The method reduces the dimensionality of the image by performing PCA as a preprocessing step. Then, morphological filters are applied to the image to extract attribute profiles, which are used as input features to train an SVM classifier.

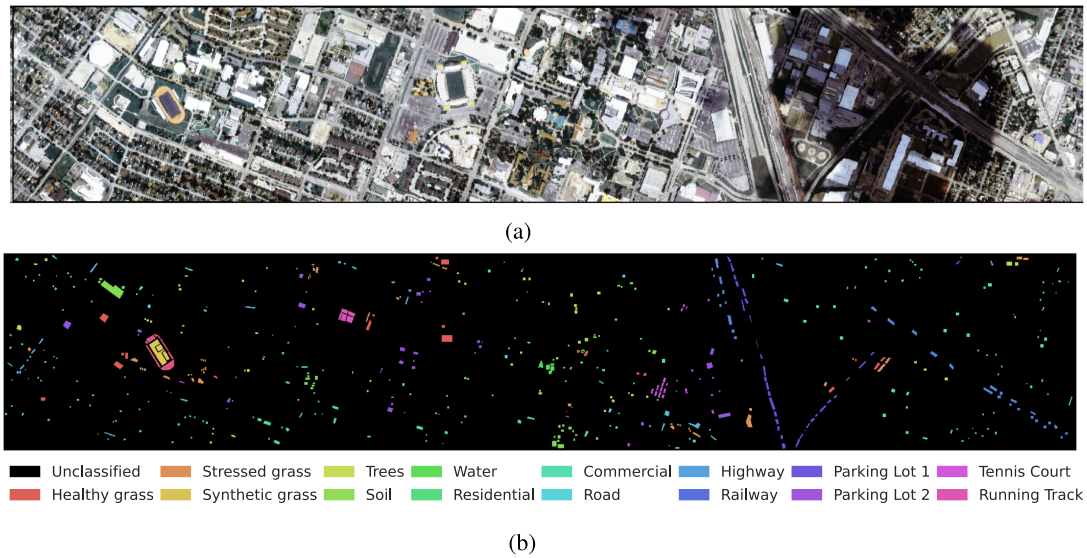


Fig. 6. Houston dataset. (a) RGB. (b) Ground truth.

TABLE IV
NUMBER OF LABELED PIXELS PER CLASS IN THE HOUSTON DATASET

No	Class Name	Number of Labeled Pixels
1	Healthy Grass	1251
2	Stressed Grass	1254
3	Synthetic Grass	697
4	Trees	1244
5	Soil	1242
6	Water	325
7	Residential	1268
8	Commercial	1244
9	Road	1252
10	Highway	1227
11	Railway	1235
12	Parking Lot 1	1233
13	Parking Lot 2	469
14	Tennis Court	428
15	Running Track	660

3) *Mini-ResNet:* A supervised baseline, i.e., training from scratch using only the labeled samples with the network architecture displayed in Fig. 2.
4) *Spectral–spatial residual network (SSRN):* A supervised baseline employing spectral and spatial residual blocks for enhanced feature extraction [10].
5) *DCFSL:* A few-shot learning approach [62] inspired from Relation Networks [30] in which the model is trained on a source scene using abundant labels and on a target scene with few annotations. To mitigate the distribution shift between source and target data, a domain adaptation loss is used.
6) *FSCF-SSL:* A state-of-the-art few-shot learning method combining meta-learning and self-supervised learning. FSCF [63] first pretrains a model following an episodic paradigm on natural images (miniImageNet), and then adapts to the target hyperspectral dataset with few samples per class. For our experiments, we follow the same setting as the original article.
7) *Barlow-Twins (BT):* Pretraining with the BT algorithm followed by linear classifier training, without utilizing pair selection (i.e., augmenting the same patch twice). This baseline aligns with previous works employing self-supervised learning for HSI classification [41].
8) *Barlow-Twins Neighbors (BT-Neighbors):* Pretraining with the BT algorithm and linear classifier training, utilizing neighboring patches in the image as positive pairs.
9) *Barlow-Twins Superpixels (BT-Superpixels):* Pretraining with the BT algorithm and linear classifier training, using positive pairs randomly sampled from the same superpixels generated in a first-stage segmentation step.

3) Network Architecture: Our primary focus is on the training strategy rather than the network architecture. Therefore, we employ a deep 2-D convolutional network with two residual blocks as our backbone. The network structure is depicted in Fig. 2. This architecture offers a good balance between simplicity, computational efficiency, and accuracy.

4) Hyperparameters: During pretraining, the projection head is a two-layers MLP with a hidden dimension of size 2048. We set the tradeoff parameter of BT to $\lambda = 0.05$. Pretraining is run for 100 epochs in all experiments. We utilize the LARS optimizer [64] with the same learning rate for all datasets, which we divide by 10 after 60 and 80 epochs. We use a batch size of 256. Our spatial neighbors pair selection strategy samples positive pairs from 9×9 neighborhoods centered around the pixel of interest. For the superpixel-based selection method, we use Felzenszwalb's efficient graph-based segmentation algorithm [58]. For classification, we adopt the linear classification protocol unless stated otherwise. A cross-entropy loss is optimized using SGD with momentum for 100 epochs. The learning rate in the classification phase is fixed per dataset. Unless explicitly stated, the patch size in all experiments is set to 9×9.

5) Data Augmentation: Unless explicitly stated, we use Gaussian noise, random flipping (horizontal and vertical) and

TABLE V
CLASS-WISE ACCURACY, OA, AA, AND KAPPA COEFFICIENT OF DIFFERENT METHODS ON THE PAVIAU DATASET USING FIVE LABELED PIXELS PER CLASS

Class	SVM	EMP-SVM	Mini-ResNet	SSRN	DCFSL	FSCF-SSL	BT	BT-Neighbors	BT-Superpix
1	58.23	87.28	79.99 ± 3.66	87.06 ± 4.64	87.00 ± 4.00	91.08 ± 3.52	89.77 ± 0.87	90.49 ± 1.48	$\mathbf{91.92 \pm 2.29}$
2	29.53	40.10	29.69 ± 5.48	37.47 ± 6.98	64.61 ± 2.34	78.79 ± 10.19	74.37 ± 2.84	80.13 ± 2.92	$\mathbf{90.64 \pm 4.91}$
3	17.86	52.01	64.51 ± 6.28	51.65 ± 13.91	57.40 ± 6.52	93.66 ± 3.24	62.87 ± 3.32	78.84 ± 4.78	$\mathbf{97.22 \pm 2.01}$
4	94.41	95.45	96.05 ± 0.98	$\mathbf{96.19 \pm 1.06}$	95.55 ± 1.29	94.26 ± 3.37	94.69 ± 1.34	92.64 ± 4.07	91.86 ± 1.88
5	99.33	99.85	99.20 ± 0.42	99.62 ± 0.17	99.82 ± 0.19	$\mathbf{99.96 \pm 0.07}$	99.86 ± 0.21	99.77 ± 0.36	99.93 ± 0.13
6	51.59	54.56	59.77 ± 2.99	54.96 ± 2.62	58.82 ± 5.46	$\mathbf{91.18 \pm 8.30}$	57.79 ± 2.00	69.78 ± 6.75	89.16 ± 6.29
7	95.70	97.28	87.05 ± 4.39	90.97 ± 2.51	78.75 ± 6.67	80.77 ± 7.19	93.12 ± 3.53	97.93 ± 1.56	$\mathbf{99.88 \pm 0.17}$
8	83.98	89.05	63.94 ± 6.56	57.56 ± 9.56	66.12 ± 11.20	92.64 ± 3.98	81.35 ± 3.25	86.53 ± 2.99	$\mathbf{97.63 \pm 1.57}$
9	99.68	99.89	98.23 ± 0.85	99.08 ± 0.75	99.26 ± 0.42	$\mathbf{99.99 \pm 0.03}$	98.06 ± 0.58	86.05 ± 3.18	86.53 ± 2.21
OA	51.12	62.84	55.90 ± 2.48	58.81 ± 3.37	71.70 ± 0.94	86.37 ± 4.98	78.20 ± 1.64	83.20 ± 1.37	$\mathbf{92.16 \pm 2.01}$
AA	70.00	79.50	75.38 ± 1.25	74.95 ± 2.57	78.59 ± 1.60	91.37 ± 1.99	83.53 ± 1.14	86.92 ± 1.12	$\mathbf{93.86 \pm 0.60}$
Kappa	42.80	54.90	47.59 ± 2.37	50.38 ± 3.48	64.42 ± 1.21	82.66 ± 6.02	72.10 ± 1.96	78.31 ± 1.73	$\mathbf{89.80 \pm 2.53}$

The best metrics are highlighted in bold.

TABLE VI
CLASS-WISE ACCURACY, OA, AA, AND KAPPA COEFFICIENT OF DIFFERENT METHODS ON THE SALINAS DATASET USING FIVE LABELED PIXELS PER CLASS

Class	SVM	EMP-SVM	Mini-ResNet	SSRN	DCFSL	FSCF-SSL	BT	BT-Neighbors	BT-Superpix
1	98.75	98.85	93.92 ± 10.30	98.72 ± 2.44	99.27 ± 0.83	99.56 ± 0.88	98.54 ± 1.51	99.12 ± 0.59	$\mathbf{99.94 \pm 0.13}$
2	74.87	96.45	93.54 ± 2.23	93.52 ± 9.66	99.51 ± 0.40	98.17 ± 2.47	95.38 ± 2.25	97.50 ± 1.21	$\mathbf{99.95 \pm 0.16}$
3	26.43	99.75	80.74 ± 9.52	65.56 ± 12.70	90.60 ± 7.98	93.59 ± 4.39	92.70 ± 1.49	96.43 ± 1.48	$\mathbf{99.98 \pm 0.06}$
4	99.57	99.71	98.68 ± 0.83	99.94 ± 0.07	99.67 ± 0.76	98.43 ± 1.96	94.92 ± 3.99	96.65 ± 3.70	$\mathbf{99.77 \pm 0.36}$
5	98.20	96.60	98.69 ± 0.91	98.16 ± 1.72	92.20 ± 3.76	99.47 ± 0.34	99.03 ± 0.66	98.36 ± 0.77	$\mathbf{99.52 \pm 0.27}$
6	95.37	96.56	98.31 ± 1.17	97.26 ± 1.65	98.73 ± 1.16	99.42 ± 0.67	98.09 ± 1.31	98.07 ± 0.97	$\mathbf{99.73 \pm 0.25}$
7	99.47	**99.64**	99.41 ± 0.27	99.77 ± 0.23	99.33 ± 0.62	99.90 ± 0.07	98.82 ± 0.65	99.03 ± 0.47	99.51 ± 0.23
8	73.32	72.29	49.15 ± 10.83	42.46 ± 12.52	72.28 ± 3.91	84.06 ± 4.13	56.94 ± 2.35	63.39 ± 3.72	$\mathbf{90.32 \pm 1.77}$
9	97.71	99.00	99.86 ± 0.18	98.77 ± 0.99	99.95 ± 0.06	99.29 ± 0.73	97.95 ± 0.22	98.99 ± 0.50	$\mathbf{99.67 \pm 0.17}$
10	11.30	89.06	91.36 ± 2.49	91.01 ± 4.23	91.43 ± 2.30	98.00 ± 1.76	80.23 ± 6.09	92.90 ± 1.81	$\mathbf{98.99 \pm 0.43}$
11	83.44	94.94	95.96 ± 1.54	94.98 ± 3.06	98.16 ± 1.45	99.58 ± 0.71	98.66 ± 0.83	98.42 ± 0.81	$\mathbf{99.05 \pm 0.22}$
12	90.53	74.09	99.81 ± 0.35	$\mathbf{100.00 \pm 0.00}$	98.90 ± 0.79	98.07 ± 0.95	74.64 ± 1.77	83.17 ± 3.65	99.99 ± 0.02
13	98.46	97.15	$\mathbf{100.00 \pm 0.00}$	99.57 ± 0.22	98.75 ± 1.39	98.70 ± 0.98	98.55 ± 0.47	99.02 ± 0.61	99.01 ± 0.85
14	90.05	96.24	98.16 ± 0.96	98.08 ± 0.52	97.97 ± 1.50	$\mathbf{99.53 \pm 0.52}$	96.55 ± 0.80	96.99 ± 2.07	85.69 ± 5.21
15	45.04	65.26	82.16 ± 7.74	88.58 ± 3.90	85.96 ± 2.13	65.98 ± 10.91	85.69 ± 1.01	88.14 ± 1.47	$\mathbf{92.75 \pm 1.64}$
16	65.82	78.91	86.84 ± 3.18	87.06 ± 5.11	91.56 ± 2.90	$\mathbf{99.15 \pm 1.05}$	83.82 ± 2.44	86.50 ± 2.15	97.64 ± 1.10
OA	74.46	86.19	84.27 ± 1.30	83.15 ± 2.86	90.46 ± 0.78	91.28 ± 1.57	85.08 ± 0.84	88.37 ± 0.75	$\mathbf{96.42 \pm 0.22}$
AA	78.00	90.90	91.67 ± 1.08	90.84 ± 1.77	94.64 ± 0.69	95.68 ± 0.84	90.65 ± 0.70	93.30 ± 0.54	$\mathbf{97.59 \pm 0.38}$
Kappa	71.50	84.60	82.64 ± 1.38	81.42 ± 3.10	89.42 ± 0.85	90.28 ± 1.75	83.49 ± 0.92	87.12 ± 0.81	$\mathbf{96.03 \pm 0.25}$

The best metrics are highlighted in bold.

random band drop as our data augmentations to generate positive pairs. The rest of the transforms are explored in Section IV-B5. In addition, we apply random resized crop to the vanilla BT baseline (without pair mining) since it is critical to learn good representations. However, random resized crop is not necessary when a pair selection strategy is used.

6) Training Maps: We consider a few-shot setting where a limited number of labeled pixels per-class K is available. Unless stated otherwise, we use $K = 5$ randomly selected pixels from the ground truth of each dataset.

7) Evaluation Metrics: We report the classification results using classical metrics, including overall accuracy (OA), average accuracy (AA), and the Kappa coefficient (κ). All experiments, including both the pretraining and classification stages, are repeated ten times with fixed training splits. We report the average value and standard deviation of each metric in the experiments.

8) Hardware: We run our experiments on servers equipped with 4 $\times$ NVIDIA A100 Tensor Core GPU and a AMD EPYC 7402 CPU.

B. Experimental Results

In this section, we present and analyze our results on the selected datasets. Tables V–VII show the class-wise accuracies, OAs, AAs, and Kappa coefficient on PaviaU, Salinas, and the WHU-Hi-HanChuan dataset. In addition, we perform several ablations to evaluate the most critical components in the proposed approach.

1) Results on PaviaU: On the PaviaU dataset, the SVM baseline performs worst due to its inability to leverage spatial information. In addition, the supervised baselines (Mini-ResNet and SSRN) also yield poor results, which is explained by the small size of the training set. Indeed, training a neural network from scratch with a limited number of labeled samples ($K = 5$ per class) leads to severe overfitting and performs worse than classical baselines such as EMP-SVM. However, we observe that pretraining significantly outperforms supervised learning in all settings, demonstrating the model's ability to leverage unlabeled pixels in the scene. Furthermore, Table V confirms that a good pair-mining strategy enriches the supervision signal during pretraining and improves the overall performance. Using neighboring patches (BT-Neighbors) as positive pairs outperforms pretraining without pair selection (BT). Moreover, employing superpixels for pair selection (BT-Superpix) achieves the best performance with an OA of 92.16% in our experiments, outperforming few-shot learning algorithms such as DCFSL and FSCF-SSL by a significant margin. The improvements are also consistent in the class-wise metrics, where significant boosts of

TABLE VII
CLASS-WISE ACCURACY, OA, AA, AND KAPPA COEFFICIENT OF DIFFERENT METHODS ON THE WHU-HI-HANCHUAN DATASET USING FIVE LABELED PIXELS PER CLASS

Class	SVM	EMP-SVM	Mini-ResNet	SSRN	DCFSL	FSCF-SSL	BT	BT-Neighbors	BT-Superpix
1	44.92	68.91	79.34 ± 4.57	78.77 ± 7.41	84.80 ± 5.61	86.54 ± 7.59	88.29 ± 3.74	91.47 ± 1.94	**95.54 ± 0.70**
2	15.54	56.52	34.41 ± 6.33	29.39 ± 8.50	38.39 ± 6.88	**92.75 ± 3.85**	58.18 ± 4.10	67.37 ± 6.23	58.59 ± 5.17
3	57.40	56.56	54.93 ± 9.54	64.29 ± 3.63	57.88 ± 5.47	52.57 ± 6.87	76.13 ± 3.14	77.83 ± 3.34	**79.28 ± 2.42**
4	66.21	81.14	87.95 ± 3.22	91.26 ± 3.33	89.98 ± 2.58	89.95 ± 10.06	79.10 ± 3.36	89.17 ± 4.14	**96.46 ± 1.20**
5	73.47	97.74	96.58 ± 1.64	96.18 ± 3.92	87.67 ± 8.13	93.97 ± 5.88	95.03 ± 2.31	98.87 ± 0.75	**99.78 ± 0.15**
6	3.29	30.31	26.46 ± 2.94	25.75 ± 2.81	22.41 ± 6.27	49.30 ± 13.06	38.39 ± 3.25	51.86 ± 4.49	**78.84 ± 4.85**
7	51.89	57.38	79.79 ± 11.12	73.42 ± 5.13	85.68 ± 3.82	68.19 ± 16.72	84.98 ± 1.28	93.29 ± 1.21	**94.20 ± 0.76**
8	6.17	33.04	25.87 ± 3.48	30.97 ± 2.42	29.50 ± 4.86	50.14 ± 10.79	37.42 ± 5.52	45.42 ± 7.17	**75.92 ± 2.98**
9	0.04	11.10	19.22 ± 9.40	34.31 ± 8.26	30.26 ± 11.02	62.22 ± 10.98	51.63 ± 5.59	**65.51 ± 6.95**	60.19 ± 5.21
10	11.75	24.90	45.37 ± 15.26	71.27 ± 10.04	63.95 ± 6.26	53.69 ± 7.93	**85.36 ± 3.92**	84.89 ± 3.40	74.68 ± 5.07
11	85.60	**93.76**	74.42 ± 11.87	71.11 ± 7.84	71.63 ± 5.98	65.03 ± 10.65	62.44 ± 3.76	63.82 ± 6.55	72.69 ± 4.57
12	25.06	52.53	41.80 ± 7.45	64.03 ± 7.40	43.23 ± 5.86	84.74 ± 6.88	80.30 ± 2.92	90.72 ± 1.63	**95.43 ± 1.27**
13	47.19	46.44	39.75 ± 3.22	48.94 ± 5.83	37.52 ± 4.78	44.10 ± 4.96	50.77 ± 2.72	47.64 ± 2.45	**55.70 ± 4.57**
14	16.92	53.42	67.20 ± 6.53	57.87 ± 5.19	67.36 ± 3.69	76.11 ± 5.87	62.68 ± 5.33	75.01 ± 4.04	**76.20 ± 3.44**
15	68.25	58.36	86.71 ± 0.70	84.62 ± 2.49	78.19 ± 7.87	86.99 ± 5.39	96.01 ± 4.47	**97.09 ± 1.53**	79.24 ± 13.05
16	70.79	70.52	92.98 ± 3.58	92.21 ± 4.58	93.88 ± 1.65	95.79 ± 2.61	92.58 ± 2.12	93.20 ± 1.12	**97.88 ± 0.39**
OA	45.25	60.25	67.30 ± 1.94	68.52 ± 2.10	70.19 ± 1.38	79.00 ± 1.58	75.32 ± 1.25	79.66 ± 0.86	**84.13 ± 1.02**
AA	40.30	55.80	59.54 ± 2.38	63.40 ± 1.18	61.40 ± 1.01	72.01 ± 1.77	71.21 ± 1.31	77.08 ± 0.77	**80.68 ± 1.55**
Kappa	37.75	54.70	62.15 ± 2.17	63.69 ± 2.29	65.45 ± 1.54	75.47 ± 1.77	71.44 ± 1.43	76.41 ± 1.01	**81.52 ± 1.19**

The best metrics are highlighted in bold.

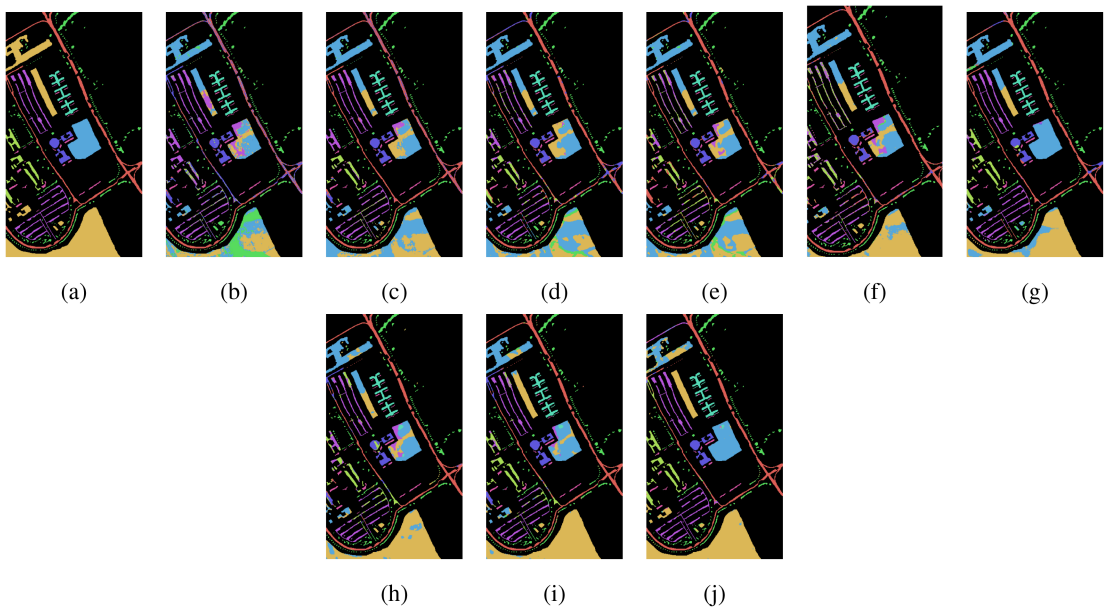


Fig. 7. Classification maps obtained on the PaviaU dataset with five labeled pixels per class. (a) Ground truth map. (b) SVM (OA 51.12%). (c) EMP-SVM (OA 62.84%). (d) Mini-ResNet (OA 58.02%). (e) SSRN (OA 58.81%). (f) DCFSL (OA 71.21%). (g) FSCF-SSL (OA 86.37%). (h) BT (OA 75.89%). (i) BT-Neighbors (OA 84.22%). (j) BT-Superpix (OA 92.10%).

accuracy are obtained on the Meadows (2), Gravel (3), and Bare Soil (6) classes. Qualitative results, shown in Fig. 7, exhibit similar trends, with BT-Superpix providing smoother classification maps and reduced confusion between classes.

To gain insight into the learned embedding space after pretraining, we project the representations of labeled patches in the dataset using t-SNE [65]. The resulting visualizations for a randomly initialized encoder, BT, and BT-Superpix are shown in Fig. 11. Despite being learned in an unsupervised fashion, the t-SNE plots indicate that samples belonging to the same classes are well clustered together after pretraining. This clustering effect becomes more pronounced with the introduction of a pair-mining strategy, particularly with BT-Superpix.

2) Results on Salinas: On the Salinas dataset, EMP-SVM is a competitive baseline, achieving 86.19% OA on average and outperforming both the supervised baselines (Mini-ResNet and SSRN) and pretraining without pair mining (BT). This demonstrates the effectiveness of traditional feature extractions techniques when labels are scarce. However, introducing pair selection mechanisms leads to significant improvements in accuracy. Indeed, BT-Neighbors, albeit its simplicity, increases the OA by more than 3% compared to BT. This highlights the importance of having good views for pretraining with contrastive learning. Most notably, the superpixel-guided sampling strategy provides a large gain in performance, reaching 96.42% of accuracy on average and significantly improving the results for

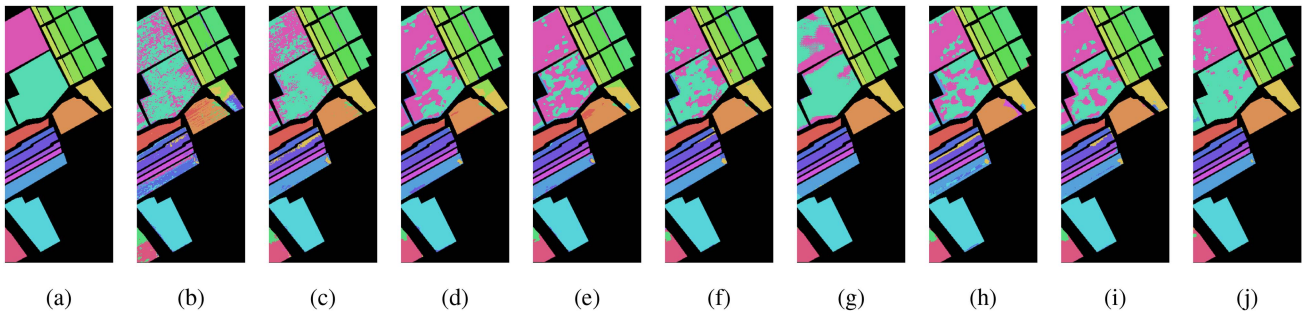


Fig. 8. Classification maps obtained on the Salinas dataset with five labeled pixels per class. (a) Ground-truth map. (b) SVM (OA 74.46%). (c) EMP-SVM (OA 86.19%). (d) Mini-ResNet (OA 83.05%). (e) SSRN (OA 83.15%). (f) DCFSL (OA 89.57%). (g) FSCF-SSL (OA 91.28%). (h) BT (OA 84.28%). (i) BT-Neighbors (OA 88.33%). (j) BT-Superpix (OA 96.45%).

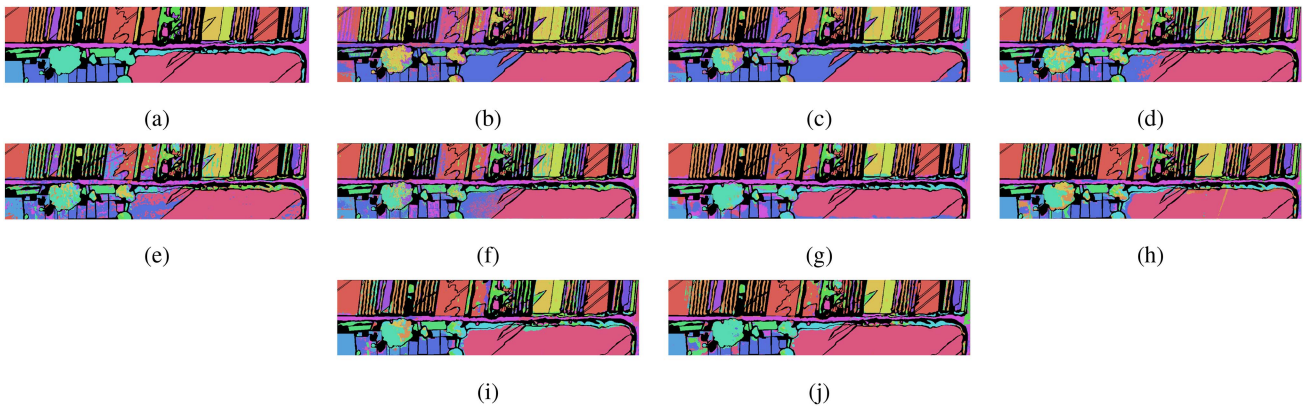


Fig. 9. Classification maps obtained on the WHU-Hi-HanChuan dataset with five labeled pixels per class. (a) Ground-truth map. (b) SVM (OA 45.25%). (c) EMP-SVM (OA 60.25%). (d) Mini-ResNet (OA 66.98%). (e) SSRN (OA 68.52%). (f) DCFSL (OA 70.21%). (g) FSCF-SSL (OA 79.00%). (h) BT (OA 77.66%). (i) BT-Neighbors (OA 79.55%). (j) BT-Superpix (OA 83.87%).

some classes such as Grapes untrained (8) and Vinyard vertical trellis (16). BT-Superpix also outperforms the few-shot learning algorithms (DCFSL and FSCF-SSL) despite their use of external data and annotations for pretraining. The resulting classification maps are depicted in Fig. 8. Inline with the quantitative results, BT-Superpix provides the closest classification maps to the ground truth.

Fig. 12 shows t-SNE plots of the embedding space learned after pretraining with and without pair selection, compared to a randomly initialized network. The plots reveal that an effective pair-mining strategy aids in clustering the classes, enabling easy separation using a lightweight classification protocol, such as a simple linear classifier.

3) Results on WHU-Hi-HanChuan: On the WHU-Hi-HanChuan dataset, the trends we observe align with the two previous datasets. The SVM baseline yields the lowest accuracy as it disregards spatial information. EMP-SVM performs significantly better thanks to its feature extraction step, but is outperformed by plain supervised learning (Mini-ResNet and SSRN) in the limited label regime. FSCF-SSL achieves competitive results thanks to its combination of meta-learning and self-supervised learning. Compared to supervised learning, vanilla BT helps mitigating overfitting and considerably improves the results compared to training from scratch. When pair-selection techniques are introduced, an increase of 4% in OA is observed with the neighbor-sampling strategy, and a gain of over 8% is achieved when selecting positive pairs using superpixels. This confirms the importance of pair selection during pretraining. Fig. 9 displays the classification maps obtained using the considered approaches. We can visually observe that BT-Superpix yields the highest quality maps.

Fig. 13 visualizes the representations learned from pretraining after a t-SNE projection. Similar to the previous datasets, the benefit of pretraining and positive pair mining in clustering the classes and reducing confusion is visually apparent.

4) Results on Houston 2013: On the Houston dataset, deep-learning-based supervised baselines (Mini-ResNet and SSRN) outperform classical algorithms by significant margins despite the limited amount of labels available. This demonstrates the effectiveness of end-to-end representation learning for classification tasks. However, pretraining using BT with and without positive pair mining outperforms the supervised baselines. Most notably, BT-Neighbors and BT-Superpix outperform the few-shot learning algorithms (DCFSL and FSCF-SSL) by approximately 12% and 4%, respectively, in OA. It is also interesting to observe that while positive pair mining helps, BT-Superpix and BT-Neighbors yield similar performance. This can be explained by the structure of the Houston scene, which has smaller and more scattered annotations, making superpixel segmentation less helpful than for the other datasets.

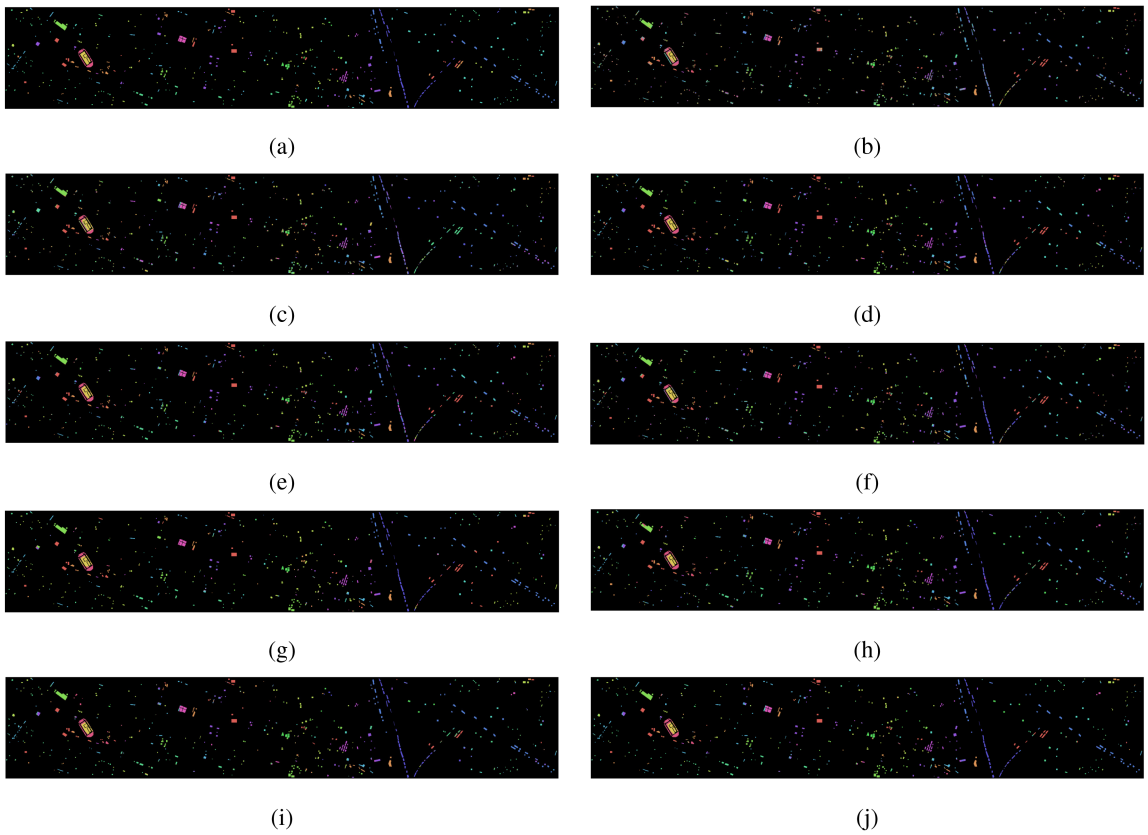


Fig. 10. Classification maps obtained on the Houston 2013 dataset with five labeled pixels per class. (a) Ground truth map. (b) SVM (OA 42.07%). (c) EMP-SVM (OA 48.77%). (d) Mini-ResNet (OA 66.48%). (e) SSRN (OA 65.47%). (f) DCFSL (OA 63.03%). (g) FSCF-SSL (OA 72.70%). (h) BT (OA 71.18%). (i) BT-Neighbors (OA 76.64%). (j) BT-Superpix (OA 77.29%).

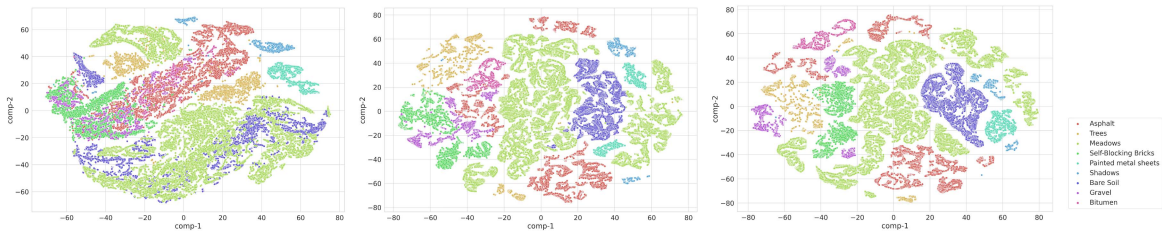


Fig. 11. t-SNE visualizations of the representations learned during pretraining on the PaviaU dataset. (a) Random Encoder (b) BT (c) BT-Superpix.

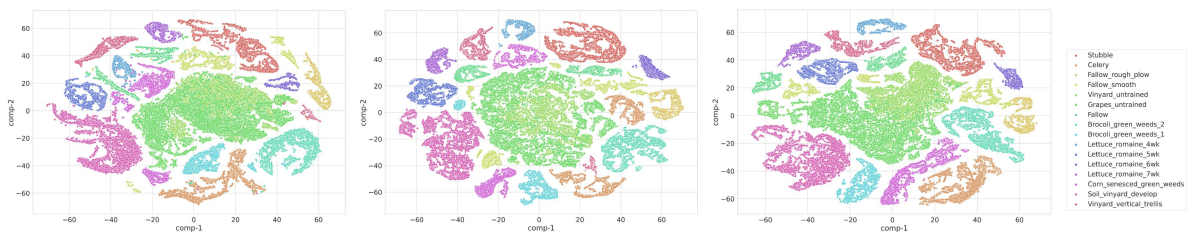
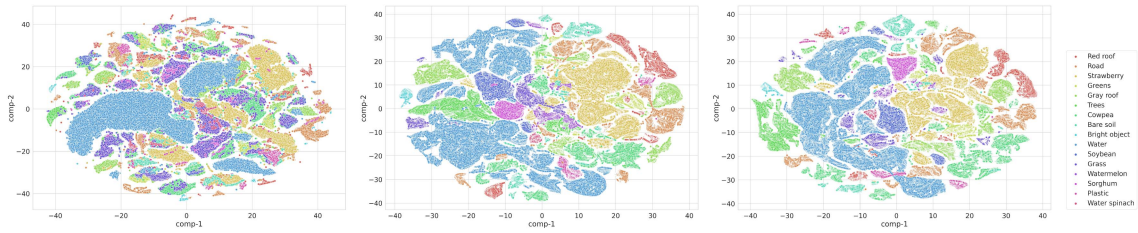


Fig. 12. t-SNE visualizations of the representations learned during pretraining on the Salinas dataset. (a) Random Encoder (b) BT (c) BT-Superpix.



(a) (b) (c)

Fig. 13. t-SNE visualizations of the representations learned during pretraining on the WHU-Hi-HanChuan dataset. (a) Random Encoder. (b) BT. (c) BT-Superpix.

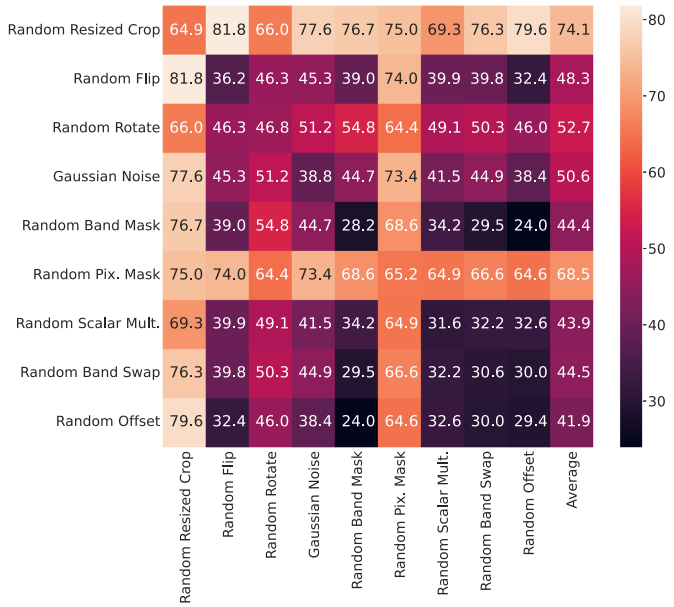


Fig. 14. Impact of data augmentation on BT evaluated on PaviaU with five labels per class for training.

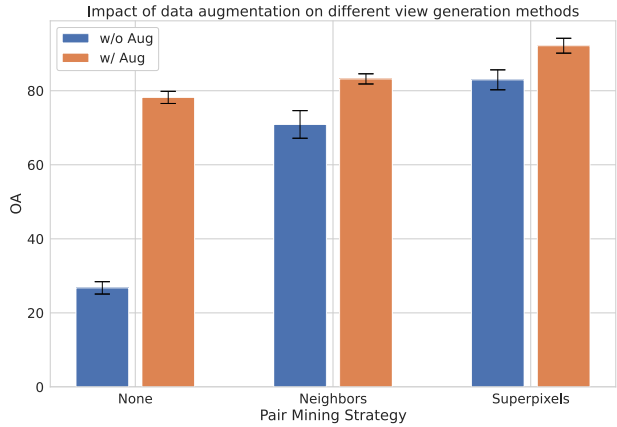


Fig. 15. Impact of removing data augmentation on different pair selection strategies evaluated on PaviaU.

5) Analysis of Data Augmentation: We analyze the impact of data augmentation on the OA of BT on the PaviaU dataset *without pair sampling*. Given the prohibitively large number of combinations of transforms, we consider up to two augmentations at a time to reduce the computational cost. We run the pretraining and classification phase for every pair of augmentations listed in Section III-C. The transforms are applied in both branches symmetrically with a fixed order and probability $p = 0.75$. Results are presented as a symmetric matrix in Fig. 14.

First, we observe a dramatic drop in performance when random resized crop is not used. This result is consistent with existing literature on computer vision [36], where the occlusion effect resulting from partial crops is essential in contrastive learning. Interestingly, the pixel removal transform also makes the pretraining task challenging enough so that the network learns meaningful representations. This finding aligns with the recent literature in masked image modeling [66], although convolutional neural networks are not well-suited for this type of masking [67].

We further observe that augmentations such as random flipping, Gaussian noise, translation, and random band dropping work well in conjunction with random cropping. The high variance of the results illustrates the importance of carefully selecting the augmentations in contrastive learning. This is a challenging problem in HSI classification given the following.

1) The spatial context in patches is very limited compared to the common setting in computer vision. In the most extreme case, random cropping is not possible with purely spectral classifiers.
2) Color distortions, which are crucial in preventing the network from relying on low-level color statistics to solve the pretraining task [36], are not easily applicable to HSIs as the relevant information lies in the spectrum of each pixel.

A data-driven approach, such as positive pair selection, offers a way to overcome these limitations. It provides a means to generate pairs that circumvent the challenges imposed by limited spatial context and the inability to heavily distort color information.

To evaluate the robustness of the proposed pair sampling strategies to the choice of data augmentations, we measure the OA on PaviaU when training BT, BT-Neighbors, and BT-Superpix *without any augmentations*. The results are presented in Fig. 15. Without pair sampling, data augmentation is compulsory, otherwise the network cannot learn meaningful representations. However, with a proper pair selection, good representations can still be learned even without data augmentation. As expected, the supervision signal obtained by sampling patches

TABLE VIII
CLASS-WISE ACCURACY, OA, AA, AND KAPPA COEFFICIENT OF DIFFERENT METHODS ON THE HOUSTON 2013 DATASET USING FIVE LABELED PIXELS PER CLASS

Class	SVM	EMP-SVM	Mini-ResNet	SSRN	DCFSL	FSCF-SSL	BT	BT-Neighbors	BT-Superpix
1	36.21	76.73	**96.61 ± 0.71**	90.94 ± 5.51	93.83 ± 1.62	75.43 ± 10.47	67.82 ± 8.28	79.48 ± 3.71	85.65 ± 2.19
2	38.19	58.45	55.79 ± 3.45	**66.73 ± 5.17**	56.40 ± 5.26	56.99 ± 8.30	52.47 ± 2.43	59.71 ± 3.58	61.46 ± 3.16
3	71.10	72.54	85.39 ± 1.67	**87.46 ± 3.50**	83.83 ± 3.10	82.82 ± 2.95	81.05 ± 0.92	81.08 ± 2.13	79.13 ± 1.28
4	46.57	37.93	79.60 ± 2.98	79.81 ± 2.44	69.98 ± 4.53	74.12 ± 4.80	78.92 ± 0.99	79.37 ± 1.25	**79.93 ± 1.47**
5	75.73	85.69	97.03 ± 0.97	96.64 ± 0.88	92.00 ± 2.41	95.42 ± 3.83	94.69 ± 1.79	97.47 ± 2.47	**98.04 ± 0.81**
6	45.62	49.06	64.00 ± 1.98	64.06 ± 1.38	60.19 ± 5.29	70.66 ± 1.80	74.53 ± 1.18	**77.28 ± 1.72**	74.25 ± 1.80
7	16.90	35.79	40.96 ± 4.23	44.07 ± 12.71	42.59 ± 7.54	53.13 ± 8.02	61.73 ± 6.00	**79.12 ± 3.61**	77.74 ± 3.86
8	30.75	13.24	24.62 ± 3.72	25.73 ± 2.45	31.34 ± 5.62	**47.18 ± 11.15**	29.95 ± 3.06	37.97 ± 3.63	34.60 ± 3.21
9	43.95	46.83	66.62 ± 3.15	65.09 ± 4.14	49.82 ± 7.64	63.20 ± 8.18	**74.42 ± 1.94**	67.78 ± 2.38	71.19 ± 4.20
10	46.89	68.49	55.43 ± 6.05	54.66 ± 15.52	60.61 ± 7.71	89.68 ± 4.94	81.64 ± 2.59	94.76 ± 1.56	**96.58 ± 2.05**
11	21.95	17.07	54.93 ± 4.93	44.93 ± 13.17	64.27 ± 5.99	86.49 ± 4.66	77.09 ± 2.11	86.96 ± 0.53	**87.20 ± 0.60**
12	29.64	15.15	64.09 ± 5.97	54.29 ± 6.43	**69.84 ± 4.34**	62.00 ± 9.50	64.89 ± 5.41	65.82 ± 5.36	68.02 ± 5.22
13	35.99	72.63	77.84 ± 3.66	77.39 ± 5.17	66.19 ± 10.51	**94.55 ± 1.50**	89.66 ± 0.67	93.32 ± 0.51	91.53 ± 1.61
14	70.45	72.58	86.74 ± 1.56	86.22 ± 1.16	81.35 ± 1.42	**100.00 ± 0.00**	87.04 ± 0.71	89.24 ± 1.27	85.37 ± 1.52
15	60.61	52.06	81.92 ± 1.18	81.13 ± 2.57	78.89 ± 4.97	75.60 ± 3.44	**90.11 ± 2.66**	86.38 ± 1.65	85.92 ± 3.95
OA	42.07	48.77	66.48 ± 0.75	65.47 ± 2.00	65.24 ± 1.56	72.70 ± 1.27	71.18 ± 0.80	76.64 ± 0.69	**77.29 ± 0.65**
AA	44.70	51.60	68.76 ± 0.69	67.94 ± 1.62	66.74 ± 1.51	75.15 ± 1.07	73.74 ± 0.69	78.38 ± 0.56	**78.43 ± 0.52**
Kappa	37.50	44.80	63.81 ± 0.80	62.75 ± 2.14	62.44 ± 1.67	70.50 ± 1.37	68.91 ± 0.86	74.79 ± 0.76	**75.48 ± 0.70**

TABLE IX
IMPACT OF THE CHOICE OF THE PRETRAINING ALGORITHM ON THE OA OF THE PROPOSED PIPELINE FOR DIFFERENT PAIR SELECTION STRATEGIES

Algorithm	No Pair Sampling	Neighbors	Superpixel
BT	78.20 ± 1.64	83.20 ± 1.37	**92.16 ± 2.01**
MoCo	75.80 ± 1.25	80.51 ± 1.59	**88.22 ± 1.31**
SwAV	79.35 ± 1.19	82.73 ± 0.48	**87.27 ± 1.02**

from the same superpixel is richer than that of spatial neighbors, explaining why BT-Superpix outperforms BT-Neighbors both with and without data augmentation. Therefore, using different samples as positives with a good selection heuristic enhances robustness to data augmentation and improves overall performance.

6) Impact of the SSL Algorithm: The proposed view generation mechanism can seamlessly be integrated in any contrastive learning algorithm. To demonstrate this, we substitute BT with two well-known self-supervised learning algorithms, namely MoCo [37] and SwAV [25]. Results are presentedanchor in Table IX. Overall, BT performs best among the three algorithms. Furthermore, incorporating pair selection consistently improves performance, with superpixel-guided sampling yielding the best results.

7) Effect of the Patch Size: The patch size is critical hyperparameter in spatial-spectral classifiers. It is well established in the HSI classification literature that the spatial-spectral classifier generally outperform purely spectral classifiers. Moreover, increasing the spatial context can enhance the overall performance, despite an oversmoothing effect. With contrastive learning, the patch size is even more crucial because the most important augmentation, i.e., random resized crop, relies on a large spatial context.

We investigate the impact of the patch size for the supervised and self-supervised methods. Specifically, we conduct experiments using patch sizes of 5, 7, 9, 11, and 13 on PaviaU. The results are depicted in Fig. 16. Consistent with the literature, we observe that the supervised baseline (Mini-ResNet) benefits from a larger patch size, yet the OA remains limited due to the scarcity of labeled examples. When pretraining without pair

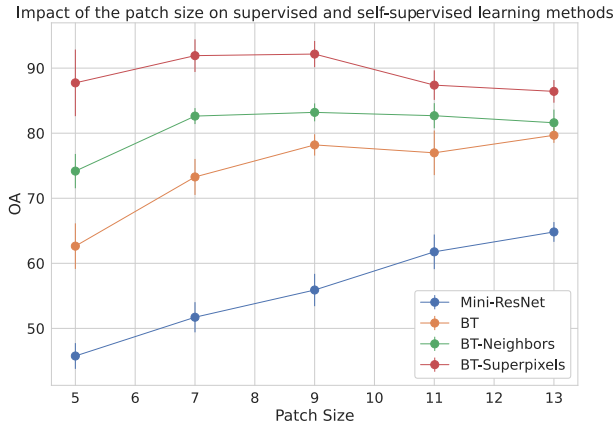


Fig. 16. Impact of the patch size on the OA of different methods on PaviaU.

sampling, we observe a substantial drop in accuracy (up to 15%) for small patch sizes. This can be partially attributed to the importance of a large spatial context during pretraining for random cropping. Moreover, the OA does not improve for values higher than 9, likely due to patches becoming less homogeneous as the spatial context becomes excessively large.

Introducing a pair selection strategy during the pretraining stage significantly enhances robustness to the patch size. In the case of BT-Neighbors and BT-Superpix, smaller spatial contexts also yield favorable results, indicating that a reduced patch size can still be competitive when pair selection is integrated into pretraining.

8) Impact of K: We focus on a few-shot setting where we have a limited number of samples per class, specifically $K = 5$ in our experiments. In this section, we investigate the impact of K on the performance of supervised and self-supervised learning methods. Specifically, we conduct experiments on PaviaU for values of K ranging between 5 and 10. The results are presented in Fig. 17.

As expected, the performance generally improves as we increase the number of labeled pixels per class. However, the accuracy curves are not strictly increasing. This behavior can

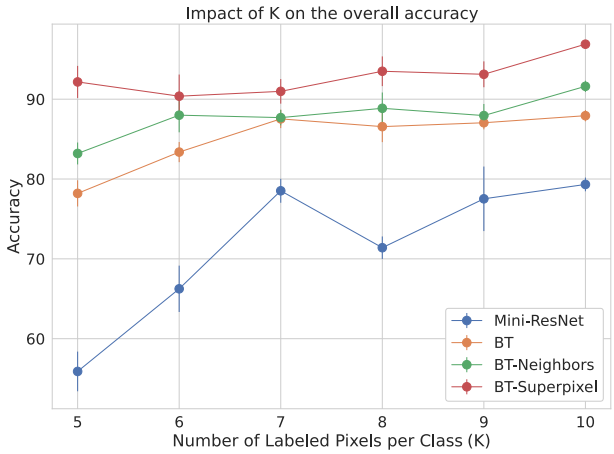


Fig. 17. Impact of K, the number of labeled pixels per class, on the OA of different methods on PaviaU.

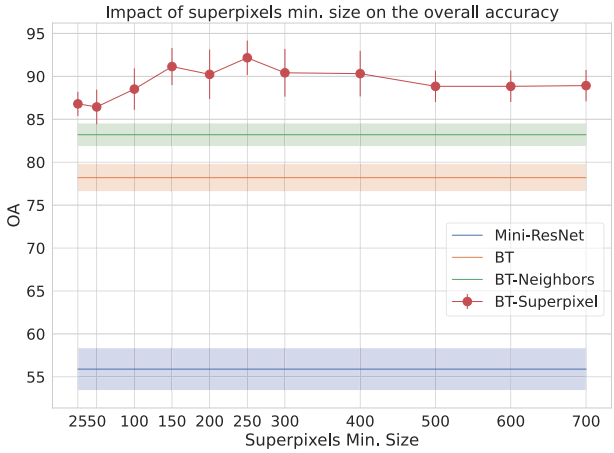


Fig. 18. Impact of the minimum size of the superpixels on the OA of BT-Superpix on PaviaU.

be attributed to the random selection of pixels in the training map, which introduces high variance in the composition of the training set. Consequently, the accuracy may fluctuate due to the specific set of pixels selected for training at each value of K.

9) Influence of the Superpixels: The effectiveness of the proposed superpixel-based pair-mining strategy depends on the quality of the superpixels generated by the algorithm. In this section, we analyze the impact of the minimum size of the superpixels, one of the most critical hyperparameters of the segmentation algorithm, on the OA of our method. The results are presented in Fig. 18.

There is a clear tradeoff between the accuracy and the variability of the positive pair selection process. When the minimal size of the superpixels is low, the resulting superpixels are smaller on average. This ensures that the positive pairs are similar and semantically coherent. However, this also reduces the variability of the training signal, thereby harming the model. This can be observed for smaller values of min size in the graph, typically below 100. In the most extreme case, where every superpixel collapses to a single pixel in the image, the method reverts back to the vanilla contrastive learning.

TABLE X
IMPACT OF THE CLASSIFICATION PROTOCOL ON THE OA OF THE PROPOSED APPROACH FOR DIFFERENT PAIR SELECTION STRATEGIES

Classification	No Pair Sampling	Neighbors	Superpixel
Linear	78.20 ± 1.64	83.20 ± 1.37	$\mathbf{92.16 \pm 2.01}$
SVM	77.16 ± 0.96	83.16 ± 1.69	$\mathbf{93.29 \pm 2.32}$
MLP	78.36 ± 1.49	83.67 ± 1.64	$\mathbf{92.58 \pm 1.99}$
Finetuning	76.55 ± 2.26	82.01 ± 1.52	$\mathbf{89.28 \pm 1.60}$

TABLE XI
COMPUTATIONAL COST OF PRETRAINING AND CLASSIFICATION ON THE PAVIAU DATASET FOR 100 EPOCHS

Stage	Training Time (in seconds)
BT	3537
BT-Neighbors	3432
BT-Superpix	3384
Linear Classifier	52

Conversely, when the superpixels are too large, the proportion of "false positives"[2] increases, thereby harming the network's performance. In the most extreme scenario, when there is only one superpixel in the entire image, the selection strategy amounts to a random sampling of positive pairs in the scene, yielding very poor representations.

To obtain the best results, it is necessary to find a suitable tradeoff between the size and the homogeneity of the superpixels. This ensures a balance between semantic coherence of the positive pairs and the variability of the training signal.

10) Influence of the Classification Protocol: After a backbone is pretrained in a self-supervised fashion, there are several ways to derive a classifier for the downstream task. In our experiments, we trained a simple linear classifier on top of the frozen backbone. In this section, we provide a justification for this design choice.

Table X presents the results obtained using four different classification protocols, including the following.

1) Linear classification with frozen backbone.
2) Training an SVM classifier with a frozen backbone.
3) Training an MLP on top of the frozen backbone.
4) Finetuning all the parameters after initializing with the pretrained model.

We observe that linear classification, despite its simplicity, consistently performed well across all pair selection strategies. The SVM and MLP classifiers only marginally improve performance. Interestingly, finetuning the backbone generally degrades performance, likely due to the small size of the training set, which makes the model prone to overfitting.

11) Computational Complexity: In this section, we delve into the computational complexity of our proposed method. We analyze the complexity in terms of the run time for pretraining and classification. We train for 100 epochs in both stages using the hardware described in Section IV-A8 with one GPU. The results for PaviaU are summarized in Table XI.

The pretraining phase is the most computationally intensive, as it processes all pixels in the HSI. The addition of positive

[2]We refer to false positives cases where two patches from different classes are drawn as a positive pair.

pair-mining strategies does not significantly increase the computational cost compared to standard contrastive learning. In contrast, the classification phase is computationally efficient, training only a single network layer on a small subset of pixels (five per class). To reduce computational costs during pretraining, one can subsample the image using a grid with a stride larger than one, thus processing fewer pixels.

V. Conclusion

In this article, we addressed the problem of few-shot HSI classification by leveraging self-supervised contrastive learning methods. We presented a two-stage pipeline that consists of the pretraining of a deep convolutional encoder using contrastive learning followed by a lightweight classification model. Our approach introduces two positive pair-mining strategies, based on spatial neighbors and superpixels, in order to generate high-quality views during pretraining. Through extensive experiments on four popular HSI classification datasets, we have demonstrated the effectiveness of our approach in improving accuracy and label efficiency. The proposed method outperforms plain supervised learning that requires larger annotated datasets and conventional contrastive learning. This highlights the potential of self-supervised learning for addressing label scarcity in few-shot HSI classification scenarios.

Furthermore, our analysis highlights the importance of data augmentation in contrastive learning and the advantages of positive pair mining. We have shown that carefully selecting positive pairs enhances the robustness and performance of the model, while reducing the need for extensive data augmentations.

In future work, we plan to investigate more advanced view generation methods, including learnable pair-mining approaches, physics-informed data augmentations, and mixing models such as ALMM [68], which would allow to better capture the inherent spectral variability and noise in HSI. Finally, we believe that generalizability of self-supervised pretraining to multisensor and cross-scene settings is an important avenue of research, which deserves further investigation.

References

[1] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[3] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[4] T.-W. Lee and T.-W. Lee, *Independent Component Analysis*. Berlin, Germany: Springer, 1998.

[5] V. Vapnik, I. Guyon, and T. Hastie, "Support vector machines," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[6] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Statistic. Math.*, vol. 44, no. 1, pp. 197–200, 1992.

[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[8] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[9] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, 2015.

[10] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[11] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[12] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[13] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[14] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[15] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[16] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[17] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[18] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self-and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82146–82168, 2021.

[19] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, 2021.

[20] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[21] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2290–2304, Apr. 2019.

[22] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[23] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[24] Y. Wang, C. Albrecht, N. A. A. Braham, L. Mou, and X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.

[25] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 9912–9924.

[26] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[27] H. Tang, Y. Li, X. Han, Q. Huang, and W. Xie, "A spatial–spectral prototypical network for hyperspectral remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 167–171, Jan. 2020.

[28] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4077–4087.

[29] K. Gao, B. Liu, X. Yu, J. Qin, P. Zhang, and X. Tan, "Deep relation network for hyperspectral image few-shot classification," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 923.

[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[31] Y. Wu, G. Mu, C. Qin, Q. Miao, W. Ma, and X. Zhang, "Semi-supervised hyperspectral image classification via spatial-regulated self-training," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 159.

[32] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Self-supervised learning with prediction of image scale and spectral order for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545715.

[33] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7758–7772, Sep. 2021.

[34] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf., Comput. Vis.*, 2020 pp. 776–794.
[35] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6008205.
[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[37] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[38] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15750–15758.
[39] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21271–21284.
[40] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5521213.
[41] X. Hu, T. Li, T. Zhou, Y. Liu, and Y. Peng, "Contrastive learning based on transformer for hyperspectral image classification," *Appl. Sci.*, vol. 11, no. 18, 2021, Art. no. 8670.
[42] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5411415.
[43] M. Zhu, J. Fan, Q. Yang, and T. Chen, "SC-EADNet: A self-supervised contrastive efficient asymmetric dilated network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5519517.
[44] J. Li, X. Li, Z. Cao, and L. Zhao, "ROBYOL: Random-occlusion-based BYOL for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 6014405.
[45] Z. Xue, B. Liu, A. Yu, X. Yu, P. Zhang, and X. Tan, "Self-supervised feature representation and few-shot land cover classification of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5541618.
[46] P. Guan and E. Y. Lam, "Cross-domain contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5528913.
[47] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 539–546.
[48] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1735–1742.
[49] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21798–21809.
[50] J. D. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=CR1XOQ0UTh-
[51] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 3967–3974.
[52] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.
[53] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.
[54] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in Earth observation," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 3, pp. 98–106, 2023.
[55] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9588–9597.
[56] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501816.
[57] C. Tao et al., "Exploring the equivalence of siamese self-supervised learning via a unified gradient framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14431–14440.
[58] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, pp. 167–181, 2004.
[59] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels," Tech. Rep., 2010.
[60] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[61] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, 2010.
[62] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5501618.
[63] Z. Li, H. Guo, Y. Chen, C. Liu, Q. Du, and Z. Fang, "Few-shot hyperspectral image classification with self-supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5517917.
[64] Y. You, I. Gitman, and B. Ginsburg, "Scaling SGD batch size to 32 k for ImageNet training," 2017, *arxiv: 1708.03888*.
[65] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[66] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[67] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, "Mcmae: Masked convolution meets masked autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 35632–35644.
[68] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Nassim Ait Ali Braham received the M.Sc. degree in computer science from Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria, in 2019, and the M.Sc. degree in artificial intelligence and data science from Université Paris Dauphine-PSL, Paris, France, in 2020. He is currently working toward the Ph.D. degree in remote sensing with the German Aerospace Center, Wessling, Germany, and with the Technical University of Munich, Munich, Germany.

In 2019, he spent six months as a Research Intern with the LIRIS-CNRS laboratory, Lyon, France. In 2020, he spent six months with the LAMSADE-CNRS laboratory, PSL Research University, Paris. In 2023, he spent six months as a Visiting Researcher with INRIA THOTH, Grenoble, France. His research interests include deep learning, computer vision, self-supervised learning, and remote sensing.

Julien Mairal received the graduate degree from Ecole Polytechnique, Palaiseau, France, in 2005, and the Ph.D. degree from Ecole Normale Superieure, Cachan, France, in 2010, both in computer science.

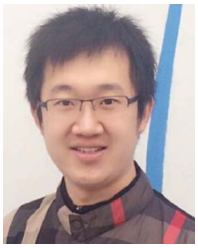
After Ph.D., he joined the Statistics Department, UC Berkeley, as a Postdoctoral Researcher. In 2012, he joined Inria, Grenoble, France, where he is currently a Research Director and Head of the Thoth team. His research interests include machine learning, computer vision, mathematical optimization, and statistical image and signal processing.

Dr. Mairal was the recipient of a starting grant and a consolidator grant from the European Research Council, respectively, in 2016 and 2022; the Cor Baayen prize in 2013; the IEEE PAMI young research award in 2017; and the test-of-time award at International Conference on Machine Learning 2019.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree in signal processing from the Université de Savoie, Annecy, France, in 1998.

From 1999 to 2023, he has been with Grenoble INP, where he was a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, USA; KTH Royal Institute of Technology, Sweden; and National University of Singapore, Singapore. Since 2013, he is an Adjunct Professor with the University of Iceland, Reykjavík, Iceland. In 2015–2017, he was a visiting Professor with the University of California, Los Angeles (UCLA). He holds the AXA chair in remote sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. He is currently a Research Director with INRIA, Grenoble. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot is a Fellow of ELLIS, a Fellow of the Asia-Pacific Artificial Intelligence Association, a Member of the Institut Universitaire de France (2012–2017), and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, since 2018). He is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007–2010), which received the 2010 IEEE GRS-S Chapter Excellence Award. He was the recipient of multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017–2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair (2009–2011) and Cochair of the GRS Data Fusion Technical Committee (2005–2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015). In 2014, he served as a Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE.



Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.-Ing. degree in remote sensing from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015, he spent six months with the Computer Vision Group, University of Freiburg, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, U.K. Since 2019, he has been an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). From 2019 to 2020, he was a Research Scientist with DLR-IMF. He is currently a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany.

Dr. Mou was the recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and a finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc. and Dr.-Ing. degrees in signal processing and remote sensing, and the "Habilitation" degree in the field of signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2020, she has been the Principal Investigator and Director of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since 2020, she has also been serving as a Director with the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, she was a Cocoordinator with the Munich Data Science Research School (www.muds.de) and the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy,; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a visiting AI Professor with ESA's Phi-lab, Frascati, Italy. She is the Chair Professor for data science in earth observation with the TUM. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., Global Urbanization, UN's SDGs, and Climate Change.

Dr. Zhu has been a Member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020–2023) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Pattern Recognition*, and served as the Area Editor responsible for special issues of IEEE SIGNAL PROCESSING MAGAZINE (2021–2023).