Self-Supervised Learning With Prediction of Image Scale and Spectral Order for Hyperspectral Image Classification

Xiaofei Yang, Weijia Cao, Yao Lu, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—In recent years, convolutional neural networks (CNNs) have achieved great success in hyperspectral image (HSI) classification attributed to their unparalleled capacity to extract the local information. However, to successfully learn the high-level semantic image features, they always require massive amounts of manually labeled data during the training process, which is expensive, scarce, and impractical, and severely hinders the improvement of supervised deep learning methods. To alleviate these burdens, we present self-supervised learning (SSL) methods for HSI classification by a pretraining model using extensive unlabeled data and fine-tuning the HSI target classification. In this article, we propose a new method for learning image characteristics by training a CNN to recognize the image scale (IS) that is applied to the HSIs. In addition, we propose a multipretext task (MT) method to learn stable and good feature representations combing two different pretext task methods and contrastive loss function. We evaluate the proposed methods in SSL benchmarks on four benchmark HSIs datasets. The experiment results demonstrate that the proposed methods outperform the traditional supervised deep learning methods when large amounts of unlabeled HSIs data are used. Moreover, it demonstrates that the SSL method is promising to alleviate dependence on manually labeled data of HSI classification. Finally, our research contributes to the creation and refinement of SSL methods for pretextual tasks within the HSIs community.**

***Index Terms*—Hyperspectral image (HSI) classification, limited labeled samples, self-supervised learning (SSL), unsupervised learning.**

I. Introduction

With the rapid development of hyperspectral sensors, hyperspectral images (HSIs) play a critical role in Earth observation missions. Unlike natural RGB images, HSIs containing more than three bands can sensitively distinguish various objects using their electromagnetic spectrum distributions. Benefiting from this advantage, HSIs are extensively utilized in a variety of fields, such as land classification [1], [2], marine monitoring [3], and urban planning [4], [5]. HSI classification is one of the most crucial topics of HSIs data processing, and it aims to identify the land cover at each pixel. Since HSIs acquisition is often disturbed by instrumental effects, and cloud noise, however, accurately identifying HSIs has attracted extensive attention.

Traditional HSI classification methods are usually divided into feature extraction and classifier. For example, the widely used support vector machine (SVM) [6] and K-nearest neighbor (KNN) [7] could be directly applied to HSI classification. The SVM can also be combined with multinomial logistic regression [8], Locality Adaptive Discriminant Analysis, and other feature extraction methods, resulting in some classification results. Recently, deep learning methods have been widely used in remote sensing, and deep learning-based HSI classification methods have gradually become a research hot topic [9], [10], [11], [12], [13]. Compared with traditional methods, deep learning-based methods can automatically capture significant features beneficial to target tasks, thus resulting in better and stable results. Particularly, convolutional neural network (CNN) (as shown in Fig. 1) is one of the most mainstream models, it can effectively extract hierarchical features from HSIs. For example, Zhang and Zhang [14] surveyed the remote sensing analysis based on artificial intelligence (AI) and pointed out that there have some challenges and opportunities for using AI in remote sensing. There have been some interesting topics of research, including the AI methods in real-world remote sensing and explainable AI algorithms in remote sensing. To fully exploit the spatial-spectral information, Yang et al. [15] proposed a 3-D-CNN for HSI classification by stacking 3-D convolution layers. The proposed 3-D-CNN achieves significant results with a large number of training samples. Different from the previous 3-D-CNN, Hamida et al. [16] replaced the pooling

Manuscript received 13 July 2022; revised 9 October 2022 and 29 October 2022; accepted 19 November 2022. Date of publication 30 November 2022; date of current version 23 December 2022. This work was supported in part by the Science and Technology Development Fund, Macau SAR (File no. 0049/2022/A1); in part by the University of Macau (File no. MYRG2022-00072-FST); in part by the Macao Young Scholars Program under Grant AM2020012; in part by the Guangdong Shenzhen Joint Youth Fund under Grant 2021A151511074; and in part by the NSFC Funds under Grant 62206073. *(Corresponding authors: Weijia Cao; Yicong Zhou.)*

Xiaofei Yang and Yicong Zhou are with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: xiaofeiyang@um.edu.mo; yicongzhou@um.edu.mo).

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the Department of Computer and Information Science, University of Macau, Macau, China, and also with Yangtze Three Gorges Technology and Economy Development Company Ltd., Beijing 101100, China (e-mail: caowj@aircas.ac.cn).

Yao Lu is with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: luyao2021@hit.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3225663

1558-0644 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

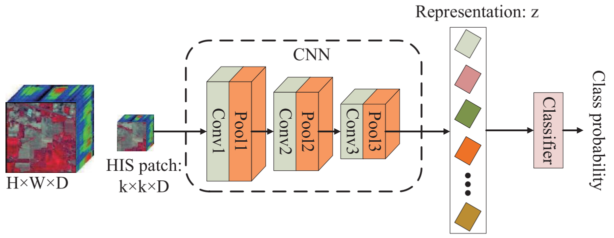


Fig. 1. CNN-based HSI classification method (supervised learning framework).

layers with the spectral-spatial 3-D convolutional layers to retrieve rich spectral-spatial features for HSI classification.

The accuracy of HSI classification has vastly increased with the implementation of diverse deep learning-based algorithms. However, all these existing deep learning-based HSI classification methods generally require a large number of labeled data. Therefore, similar to ImageNet [17] containing more than ten million labeled RGB images, it is urgent to build the same-scale HSIs dataset. But it is severely expensive and impossible to manually label such a massive HSIs dataset using extensive expertise and regional knowledge. Some researchers first pretrained the CNNs using a huge natural image dataset such as ImageNet [17] or CIFAR-10 [18]. The pretrained CNNs were fine-tuned using the target small remote sensing dataset [4], [19], [20]. However, the fine-tuning step cannot be applied immediately to HSIs because of the significant difference between HSIs and RGB images. There are some semisupervised learning methods for HSI classification tasks. For example, Zhan et al. [21] proposed a novel semisupervised method based on the 1-D generative adversarial network (GAN) for HSI classification. Zhong et al. [22] proposed a semisupervised GAN for HSI classification by integrating the GAN and conditional random field (CRF) to alleviate the shortage of labeled data. In these two works, the discriminators are first trained with the fake and real data, and then the pertained discriminators are fine-tuned to the target task by replacing the last layer of the discriminators with a new softmax. The proposed methods are based on GAN, however, it sometimes does not converge, and the results are unstable.

Self-supervised learning (SSL) techniques have recently attracted much significant attention from researchers for their ability to learn the model without requiring human cost samples [23], [24], [25], [26]. By handling the pretext tasks, they can learn a good feature representation from a large unlabeled source dataset. Following that, they transfer and fine-tune the pretrained models to the target tasks with fewer labeled examples. Inspired by these, the SSL technique could work for HSI classification when fewer labeled samples are used. There might be two reasons for this: 1) CNNs can be pretrained with any type of HSIs dataset without manually labeled samples and 2) due to the similar source and target data, pretraining CNNs on these data using SSL can potentially alleviate the domain difference. Therefore, to solve the little labeled samples and leverage considerably more spatial and spectral information, an image scale (IS) pretext task method is proposed in this article for HSI classification tasks. Specifically, we propose to learn a stable representation by predicting the IS and Spectral Order. The learned representation could be further fine-tuned for the target HSIs task to achieve the final HSI classification results. In addition, a multipretext task (MT) method is proposed to learn a stable and good feature representation. The contributions of this article are summarized as follows.

1) We introduce the SSL methods into HSI classification to address the problem resulting from insufficient manually labeled data in the deep learning-based methods. It is a new way to learn the feature representation from unlabeled samples.
2) We propose a new and simple self-supervised task called IS that makes full use of the spatial and spectral information, resulting in offering a powerful supervisory signal for semantic feature learning for HSI classification task.
3) We further investigate an MT method to simultaneously estimate the central pixel, boundary information and pixel rotation by combining two SSL pretext task methods, including the Image Rotation and IS methods. Additionally, we also integrate the contrastive loss function in the proposed MT. It enables the development of a stable and good feature representation.
4) Extensive experiments on four benchmark HSIs datasets demonstrate that SSL approaches for the pretext task, particularly the proposed IS method and MT method, outperform the standard CNN-based methods, pretrained methods, and other SSL methods.

The remainder of this article is organized as follows. Section II discusses related work. Section III illustrates the proposed methods and discusses SSL pretext tasks methods. Section IV contains the results of the experiments and analyses. Section V concludes with findings and recommendations of future research.

II. Related Work

A. HSI Classification Using Deep Learning-Based Methods

Benefiting from the powerful feature extraction capabilities of deep learning models, many deep learning-based HSI classification methods have been proposed for identifying HSIs [27], [28], [29], [30], [31]. For example, to capture the spatial-spectral contexture information, Yang et al. [15] proposed a 2-D-CNN model by stacking 2-D convolutional layers for HSI classification. The proposed 2-D-CNN achieves satisfactory classification results; however, it requires a lot of computing resources. To reduce computing resources, Ribalta Lorenzo et al. [32] first utilized an attention mechanism to choose bands, and then proposed a 2-D-CNN model for HSI classification. However, these deep learning-based approaches analyze spatial and spectral data separately, rather than extracting features by merging them. It will be unable to perform a satisfactory performance as a result of insufficient retrieved information.

Now, more and more 3-D-CNNs-based HSI classification methods have been proposed, which are attributed to

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

their power capability of representing spatial-spectral fusion information [16], [33], [34]. For example, Chen et al. [35] utilized the 3-D convolution layers to realize a deep 3-D-CNN model for HSI classification. To fully exploit the 2-D and 3-D convolution layers, Yang et al. [10] presented an integrated network for HSI classification by combining 2-D-CNN and 3-D-CNN into a single framework. Although these 3-D-CNNs-based algorithms have attained high performance, they require a large number of labeled samples. Obtaining a large number of accurate human-annotation tagged samples, however, is a time-consuming process that requires substantial skill and in-depth regional knowledge.

B. SSL Methods

Different from supervised learning methods, SSL approaches first acquire useful feature representations from their supervised data rather than through human annotation, and then fine-tune them to the target tasks [36], [37]. In short, SSL algorithms create their own supervised knowledge without relying on human annotation samples [23], [38], [39]. SSL methods could be divided into two categories: pretext tasks and contrastive representation. Pretext tasks aim to pretrain the methods using a pretext task, and then fine-tune the target task. For example, Misra and Maaten [37] and Doersch et al. [40] demonstrated SSL approaches for exploring spatial relationships to develop feature representations. Pathak et al. [41] also developed a novel SSL technique, e.g., image inpainting (IP), for learning a satisfactory feature representation through the use of an encoder–decoder architecture. Only the pretrained encoder was used to fine-tune the target tasks in IP. Dinh et al. [42] trained an image reconstruction model beforehand and then fine-tuned it for the target task. Several researchers choose to demonstrate SSL approaches using data enhancements, such as color modification [43] and spatial rotation [44], [45]. Additionally, there are some SSL tasks, such as image colorization [43], [46] and patch reordering [47], [48], [49], [50]. However, all of these strategies are provided for natural image problems and typically require anticipating some covariant low-level property of an image change. Except for a few super-pixel-based SSL approaches, there are few methods for HSI classification that use SSL [51]. Wang et al. [51] first introduced the SSL method to HSI classification using the CRF embedding. In this article, we propose a simple SSL method based on IS prediction, which could be used to learn invariant feature representations of image transformations rather than covariant feature representations.

Compared to pretext tasks, contrastive representation learning aims to learn an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart. For example, Chen et al. [24] proposed a simple framework for contrastive learning called SIMCLR, which defines "positive" and "negative" sample pairs that are treated differently in the loss function. It makes the representation of the positive sample similar, and the representation of the negative sample and the positive sample far away using a contrast loss. However, the negative samples are related to the batch size, which is limited by the memory of the GPU. In addition, He et al. utilized the asymmetric learning updates, and proposed the MoCo V1 [52] and MoCo V2 [53]. In these two methods, the momentum encoders and the main networks are updated separately. Another is the cluster feature representation method that uses two shared parameters in end-to-end networks to obtain different representations followed by a clustering algorithm to cluster similar sample representations. For example, k-means in Deep Cluster (DEEPCLUSTER) [54] or nondifferentiable operators in swapping assignments between multiple views (SWAV) [55]. In another recent line of work, bootstrap your own latent (BYOL) [56] and Simple Siamese (SIMSIAM) [57], both the network architecture and parameter updates are modified to introduce asymmetry. Inspired by these works, Hou et al. [58] first introduced contrastive learning to HSI classification and achieved improved results.

III. Methodology

A. Overview of SSL Architecture

In this article, we suggest using the SSL framework to solve the HSI classification task, and propose two new pretext tasks (e.g., IS and MT) for HSI classification. As shown in Fig. 2, the SSL framework is divided into two stages: 1) pretraining stage: training a CNN model using the pretext task methods (e.g., IS and MT) and 2) fine-tuning stage: fine-tune the CNN model with the target HSI classification task. It is noted that the CNN (as shown in Fig. 1) contains three convolutional operations, followed by the Max-pooling and Rectified Linear Unit (ReLU) functions. In the pretrained stage, the CNN model is trained to acquire a meaningful feature representation using a preset pretext task method based on a large number of unlabeled images. The learned function representation will be preserved in the parameters of the CNN model. In the fine-tuning stage, the pretrained model will be fine-tuned on the target task to deliver a superior outcome.

B. Proposed Pretext Task Method: IS

The purpose of SSL pretext task method for HSI classification is to train a CNN-based feature representation in the absence of target labels. To achieve this objective, we propose a unique and easy pretext task method of distinguishing cropped images to train the CNN model $F(\cdot)$. In particular, we first flip the Spectral Order of the input image to generate two images I_1 and I_2 with label z. We then create a set of K geometric transformations $G = g(\cdot \mid y)_{y=1}^{K}$, where $g(\cdot \mid y)$ crops the image $X \in (I_1, I_1)$ and gives the cropped image $X^{(z,y)} = g(X \mid (z, y))$ with varied size and label (z, y). It is noted that the central pixel is always referred to as (z, y) in the operation of the geometric transformation. The difference between $X^{(z,y)}$ is the border when using the inter-nearest function. Then, the CNN model $F(\cdot)$ accepts an input image $X^{(z*,y*)}$ (with the mask $(z*, y*)$ unknown to model $F(\cdot)$) and creates a probability group for the input image $X^{(z*,y*)}$

$$F\left(X^{(z*,y*)} \mid (W, b)\right) = F^{y}\left(X^{(z*,y*)} \mid (W, b)\right) \tag{1}$$

where $F^{y}(X^{(z*,y*)} \mid (W, b))$ is the predicted probability for the geometric transformation with label (z, y) and the (W, b) is the parameter of the CNN model.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

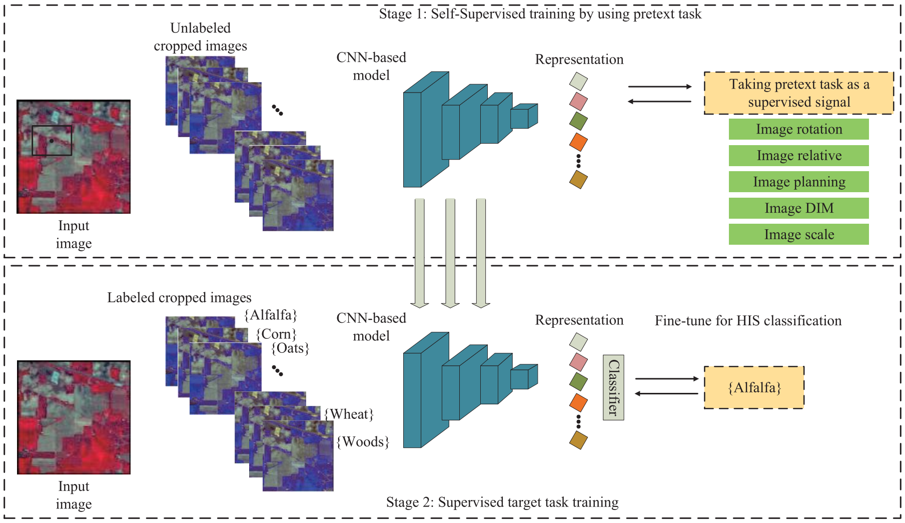


Fig. 2. SSL framework for HSI classification.

Given a set of N training samples $X_{i=0}^{N}$, the object function of the proposed IS is

$$\min \frac{1}{N}\sum_{i=1}^{N} \operatorname{loss}(X_i, (W, b)). \tag{2}$$

There are two loss function, i.e., loss_z and loss_y. The loss_z and loss_y are used to calculate the distance difference between the prediction and ground truth for the Spectral Order and IS, respectively. Both of them are calculated using the CE function. Thus, the total loss function is formulated as follows:

$$\begin{aligned} \text{loss}_{\text{total}} &= \text{loss}_z + \text{loss}_y \\ \text{loss}_z &= -\frac{1}{N}\sum_{1}^{N}\sum_{i=1}^{M} z_i \cdot \log(p_i) \\ \text{loss}_y &= -\frac{1}{N}\sum_{1}^{N}\sum_{j=1}^{C} y_j \cdot \log(q_j) \end{aligned} \tag{3}$$

where N is the total number of samples, M and C denote the number of categories of IS and Spectral Orders, respectively. z_i is the true label for category i of the Spectral Orders, while p_i represents the calculated probability value of category. y_j and q_j are the true label and the calculated probability value of the category j.

1) IS for HSI Classification: According to the above illustrations, the proposed IS should force the CNN model $F(\cdot)$ to execute an effective feature representation for HSI classification. As a result, the set of geometric transformations G is defined as all the ISs of different sizes, shown in Fig. 3. More precisely, in our work we first flip the input image I, resulting in two flipped images I_1 and I_2 with labels $z = (1, 2)$; we then recommend using G with different odd multiple sizes on these two images, e.g., 1×1, $3 \times 3, \ldots,$ and 15×15, to create $K = 8$ images. Thus, we can obtain 16 images that have the label (z, y), wherein $z \in (1, 2)$ and $y \in (1, 2, \ldots, 8)$. We note that all the generated images would be resized to the max size, such as 15×15 in our work.

a) Forcing the learning of semantic features: Learning a suitable feature representation, a good CNN model may effectively execute the aforementioned IS identification challenge. The goal of implementing IS as a sequence of geometric changes is to train the CNN model to distinguish and detect objects in images as well as their semantic parts. To be more specific, the CNN model should learn to localize salient objects in an image and distinguish their spatial and spectral order differences, object category, and boundary information. And it should effectively distinguish the scale classes of the input images using the difference information and boundary information. Fig. 4 shows some attention maps (with 128×128 patches) developed on target classification task (supervised learning) and pretext task method, e.g., IS (SSL). It is noted that the brighter part represents the focus of the model. Fig. 4(b) depicts the attention maps generated by a CNN model trained on IS, while Fig. 4(c) illustrates the attention maps created by a CNN model supervised and trained on the target classification task. These attention maps are calculated based on the activation amplitude of each spatial unit of the convolutional layer, which essentially reflects where the CNN network focuses most of its attention to classify the input image. From Fig. 4(b), we can observe that in order for the CNN model to accomplish the scale prediction task, it learns to focus on the high-level object parts surrounding the center pixel, such as the buildings, lands, spectral order, and boundary information. By comparing them with the attention maps generated by a CNN model trained on the target classification task in a supervised way [see Fig. 4(c)], we can observe that both learn a feature representation by working on approximately the same regions of the input image. Remarkably, the proposed IS method has a greater amount of variety even than the supervised method, such as the attention maps of the second and fourth rows.

b) Easy-to-detect low-level visual features: An important advantage of using IS over other pretext task methods is that they can be realized by flipping the spectral order and intercepting different sizes (as shown in Fig. 3). All the

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

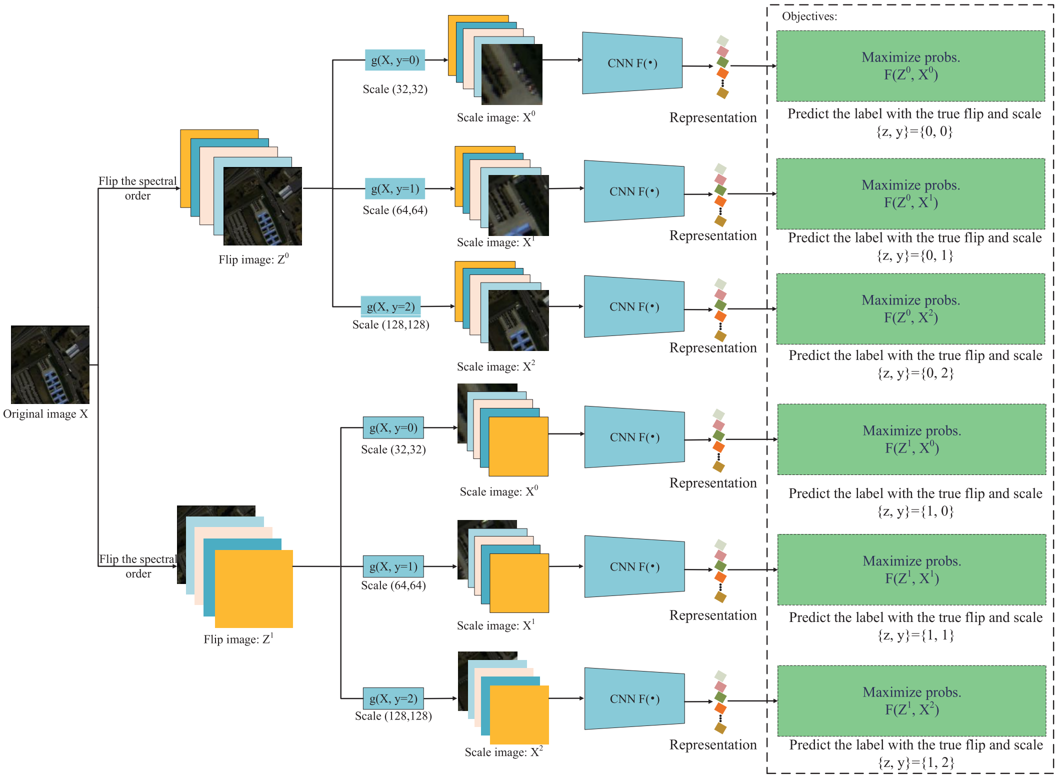

Objectives:
Original image X
Flip the spectral order
Flip image: Z^0
Flip the spectral order
Flip image: Z^1
g(X, y=0)
Scale (32,32)
g(X, y=1)
Scale (64,64)
g(X, y=2)
Scale (128,128)
Scale image: X^0
Scale image: X^1
Scale image: X^2
CNN F(•)
Representation
Maximize probs. F(Z^0, X^0)
Predict the label with the true flip and scale {z, y}={0, 0}
Maximize probs. F(Z^0, X^1)
Predict the label with the true flip and scale {z, y}={0, 1}
Maximize probs. F(Z^0, X^2)
Predict the label with the true flip and scale {z, y}={0, 2}
Maximize probs. F(Z^1, X^0)
Predict the label with the true flip and scale {z, y}={1, 0}
Maximize probs. F(Z^1, X^1)
Predict the label with the true flip and scale {z, y}={1, 1}
Maximize probs. F(Z^1, X^2)
Predict the label with the true flip and scale {z, y}={1, 2}

Fig. 3. Illustration of the proposed IS method. It first flips the input image, and then generates different scale images. The flipped and scaled images are resized the same size and then fed into the CNN architecture to learn a feature representation by predicting the flip and scale labels.

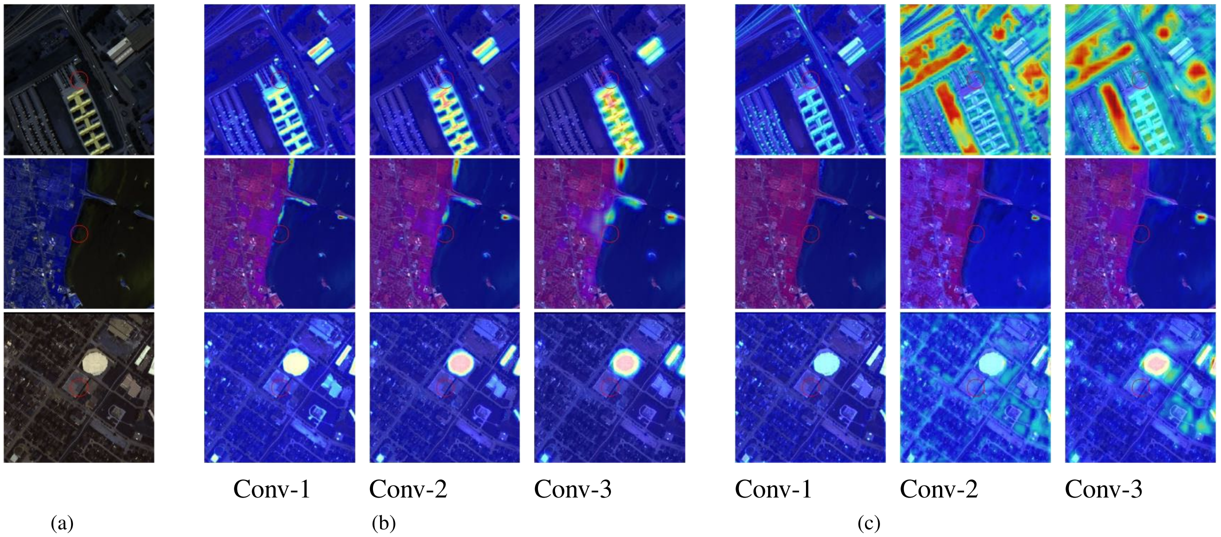


Fig. 4. Attention maps generated by the CNNs model trained on (b) IS and (c) supervised method. The brighter part represents the focus of the model. Conv-1, Conv-2, and Conv-3 are the first, second, and third Convolution layers in the CNN model, respectively. (a) Inputs. (b) Attention maps of IS. (c) Attention maps of supervised method.

generated images would be resized the same size that will lead the boundary information more and more different as the size increases, and more and more information surrounding the center pixel. In addition, the flipping operation in the spectral dimension makes the CNN model easy to learn the subtle spectral difference.

c) Well-defined: Furthermore, HSI classification is a pixel recognition task, thus making the IS task well defined. Given the input patch 15 × 15, we can get a set of different scale images by cropping different scale center pixels. And then, the generated images (such as 1 × 1, 5 × 5, and 13 × 13) are resized to 15 × 15 using the interpolation operation. With the interpolation operation, the generated patches will obtain two kinds of information, including the central pixels that are retained from the input patch 15 × 15 and the boundary pixels are obtained by interpolating from the neighborhood pixels. Thus, the generated patches are different in the boundary information and can help to learn a

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

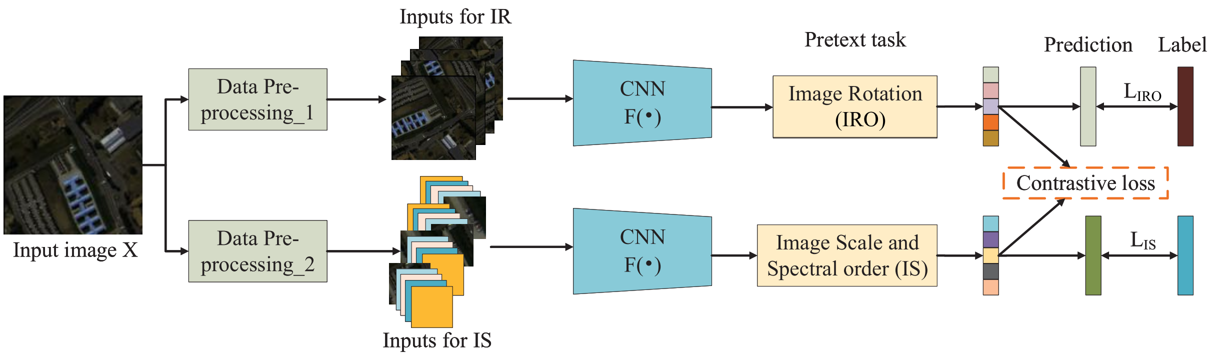


Fig. 5. Architecture of the multipretext task method by combing the IRO and IS pretext tasks. The L_{IRO} and L_{IS} denote the loss in the two pretext tasks.

good feature representation. It is noted that the central pixels of all the generated patches are equal to the central pixels of the raw input patch.

2) Discussion: Since HSI classification is equivalent to pixel classification, it is challenging to depict HSIs using Image Rotations. This displays features in a "up-standing" state in human-captured images. A significant benefit of utilizing IS over other pretext task methods, such as Image Rotation, and IP, is that a CNN model learns a feature representation by distinguishing the differences among the pixels of the input image, especially for the boundary information. In addition, it has the same computational cost as supervised learning, similar to training convergence speed. Furthermore, it can be easily embedded in other SSL models, such as Image Rotation. Despite the simplicity of our proposed method, as we will see in Section IV, the features learned by our approach have made significant improvements to the unsupervised feature learning benchmarks.

C. Multipretext Task SSL Architecture

Various pretext task SSL architectures learn the optimum feature representation for HSIs classification by gaining diverse feature attention. For example, the Image Rotation architecture focuses on the difference between the rotated pixels, while the image relative (IRE) architecture focuses on the difference between two neighboring patches. Distinct pretext task methods learn different information, and thus results in various representations. Thus, we propose an MT method with a common trunk starting with a CNN model and a head for each pretext task. Additionally, the proposed MT method also introduces the contrastive loss function.

As shown in Fig. 5, the proposed MT method is built with two pretext tasks, including the Image Rotation (IRO) and IS. Each task in the proposed MT method employs a unique separated loss function and an additional layer for a head. To construct this MT method, we first fed all inputs into the data-preprocessing modules to generate the new inputs for multiple pretext tasks (i.e., IRO and IS). It is noted that two data-preprocessing modules are used to generate two kinds of inputs for multiple pretext tasks (i.e., IRO and IS). We then put these all generated inputs into the trunk to train, and only one task was active at each training iteration. However, after several training iterations with all active pretext task, the average gradients can be computed. Inspired by Tian et al. [59], we also introduced a contrastive loss function [i.e., noise contrastive estimation (NCE) loss function] to enhance the feature learning ability. The NCE could enhance the feature learning ability of each task (i.e., IRO and IS). Hence, the total loss is the summation of all the losses from the various pretext task methods. It is noted that different loss functions are used for different tasks corresponding to the settings of their papers.

From Fig. 5, it can be observed that it is very simple to construct an MT framework based on two pretext task baseline methods, and the final loss function can be calculated by

$$\begin{aligned}\mathrm{loss} &= \mathrm{loss}_{\mathrm{pretexts}} + \alpha \mathrm{loss}_{\mathrm{contrast}} \\ &= \mathrm{loss}_{\mathrm{IRO}} + \mathrm{loss}_{\mathrm{IS}} + \alpha \mathrm{loss}_{\mathrm{contrast}} \end{aligned} \tag{4}$$

where α is a weight to balance losses. Here, the α is set as 0.5. $\mathrm{loss}_{\mathrm{IS}}$ is shown in (3). $\mathrm{loss}_{\mathrm{IRO}}$ is represented as follows:

$$\begin{aligned}\mathrm{loss}_{\mathrm{IRO}} &= \frac{1}{l} CE\left(y_i, \check{y}_i^{(w,b)}\right) \\ &= -\frac{1}{l}\sum_{i}^{K} y_i \log \check{y}_i^{(w,b)}\Big) \end{aligned} \tag{5}$$

where w and b are the parameters of the CNNs model $F(\cdot)$, $\check{y}_i$ is the projected label of input image X^i, and y_i is the true label. Finally, l is the total number of training samples.

For $\mathrm{loss}_{\mathrm{contrast}}$, we rewrite it here. It is noted that the contrastive is similar to [59]. Suppose t is the target sample, which is $P(t|C=1;\theta) = p\_(\theta)(t)$. And $\check{t}$ is the noise sample, where $P(\check{t}|C=0;\theta) = q(\check{t})$. Thus, we aim to model the logit of a sample v from the target data distribution instead of the noise distribution

$$l_\theta = \log \frac{p\_(\theta)(v)}{q(v)} = \log p\_(\theta)(v) - \log q(v). \tag{6}$$

We can convert the logits into probabilities using sigmoid $\sigma(\cdot)$, and then apply cross entropy loss

$$\begin{aligned} L_{\mathrm{NCE}} &= -\frac{1}{N}\sum_{i=1}^{N}\big[\log \sigma\left(l_\theta(t_i)\right) \\ &\quad + \log\left(1 - \sigma\left(l_\theta(\check{t}_i)\right)\right)\big] \\ \sigma(l) &= \frac{1}{1+\exp(-l)} = \frac{p\_(\theta)}{p\_(\theta)+q} \\ \mathrm{loss}_{\mathrm{contrast}} &= L_{\mathrm{NCE}}^{u^1} + L_{\mathrm{NCE}}^{u^2} \end{aligned} \tag{7}$$

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

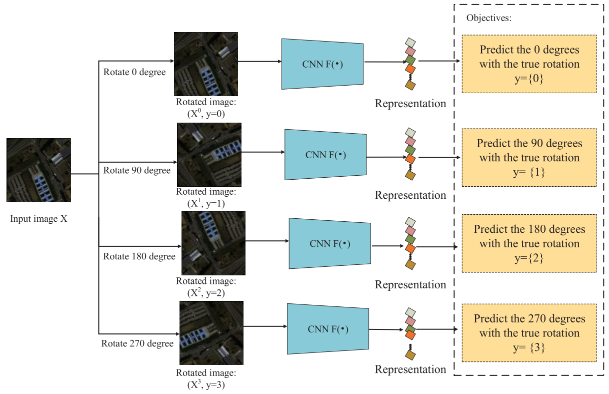

Input image X
Rotate 0 degree
Rotate 90 degree
Rotate 180 degree
Rotate 270 degree
Rotated image: (X⁰, y=0)
Rotated image: (X¹, y=1)
Rotated image: (X², y=2)
Rotated image: (X³, y=3)
CNN F(•)
Representation
Objectives:
Predict the 0 degrees with the true rotation y={0}
Predict the 90 degrees with the true rotation y= {1}
Predict the 180 degrees with the true rotation y={2}
Predict the 270 degrees with the true rotation y= {3}

Fig. 6. Illustration of the Image Rotation pretext task method. The model is used to predict the labels of the input rotated images.

where N is the number categories of the target sample. u_1 and u_2 are the output feature from IRO and IS, respectively.

D. Pretext Task SSL Architecture Implementation Details

Because each SSL methods requires distinct preprocessing of its input data, we give additional information on the heads of our pretext SSL tasks in this section. As we all know, gradients may vanish with the increasing of networks' depth, resulting in the failure of the networks. Therefore, for pretext SSL tasks, we restart training a CNN with the regular depth, which has three convolutional layers.

1) Image Scale: We first train the CNN with a set of patches. The CNN includes three convolutional layers to produce the output features, each of which contains more than 100 filters with the kernel size of 3 × 3. Each convolutional block is followed by batch-normalization and down-sampling algorithms. As a result, we may acquire the final outputs to identify the category. To be more specific, we first crop the input images to generate many new images with different scale sizes (e.g., 1×1, $3 \times 3, \ldots,$, and 15×15). We then resize all these images to the max size (such as 15×15 in our work), which are fed into the CNN model to extract the features. Finally, we add three additional fully connected layers to provide the final feature outputs. The first layer has 1024 output channels, and the second layer has eight output channels to produce the softmax outputs for predicting the input scale, and the third layer has two output channels to predict the input Spectral Order.

2) Image Rotation (IRO) [44]: We execute the CNN with four convolutional operations, just like we did with IS. To minimize the computation complex of CNNs, the patch of input image is also set as 15×15, to generate the output features. To get a satisfactory performance, the head of this task should contain more parameters. Thus, we apply two more fully connected layers to produce the softmax outputs. The first layer generates 1024 feature maps, while the second layer generates eight softmax outputs. Its architecture is illustrated in Fig. 6.

3) Image Inpainting [41]: We run the CNN including four convolutional operations and three down-sampling operations in the IP pretext task method. We first crop the input images to 15×15 and mask the central regions of the original images. The masked images are then fed to the CNN model to learn a feature representation. Using a U-Net architecture, these features to create a new image with the same size as the original image. Finally, the distance between the masked central regional and the newly generated central regional is calculated. It should be noticed that we simply preserve the parameters of the CNN model. Its framework is shown in Fig. 7.

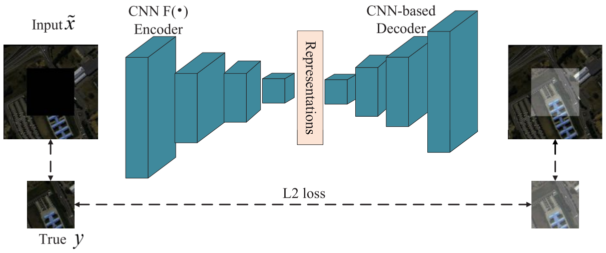


Fig. 7. Illustration of the IP pretext task method. The model is utilized to predict the labels of the input images.

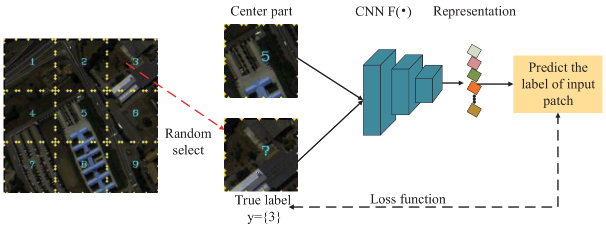


Fig. 8. Illustration of the IRE pretext task method. The model is used to predict the labels of the input images.

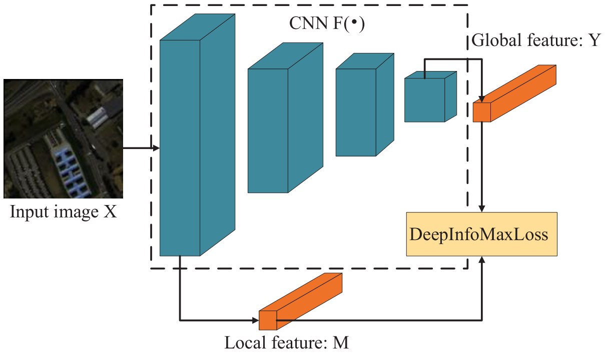


Fig. 9. Illustration of the Image DIM pretext task method. The model is used to predict the labels of the input images.

4) Image Relative [50]: The same CNN including four convolutional blocks and three max-pooling blocks are used to retrieve the IRE. The images are first cropped to 15×15, and then sampled to 3×3. The example patches are then resized to 15×15 to produce output feature. In terms of IRE, we use two fully connected layers with 1024 output channels and eight output softmax channels, respectively. Its architecture is shown in Fig. 8.

5) Image Deep InfoMax [60]: We execute the CNN, which contains four convolutional layers and three down-sampling layers for the Image DIM stage. The input images containing 15×15 are fed into the CNNs, and the distance between the

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SETS OF FOUR HSIS DATASETS IN THE EXPERIMENTS

Dataset	IN	PU	KSC	Houston
classes	16	9	13	15
unlabeled samples for SSL	10249	42776	5211	15029
labeled samples for fine-tune	5/class	5/class	5/class	5/class
samples used for testing	10169	42731	5146	14954

first and final convolutional output features is calculated. As a result, a feature representation can be learned. The architecture of DIM is shown in Fig. 9.

IV. EXPERIMENT

In this section, we provide a standardized and fair environment for evaluating our proposed methods by implementing them on four benchmark HSIs datasets. All the methods are implemented on Pytorch platform and use the same backbones, training epochs. And, the methods are evaluated using two well-known evaluation metrics [overall accuracy (OA), and Kappa (κ)]. We spread it on a desktop PC with an Intel Core 7 Duo CPU (at 3.40 GHz), 64 GB of RAM, and one GTX R3090 GPU (24 GB of ROM). We train all models with various techniques on the same computing environment and store parameters of the CNNs with the different self-supervised tasks. These stored parameters then initialize the parameters of the CNNs model for target HSI classification task.

A. Datasets Description and Experiment Designing

The proposed methods will be performed on four benchmark HSIs datasets, including Indian Pines Scene (IN), Pavia University scene (PU), Kennedy Space Center (KSC), and Houston 2013 datasets, shown in Table I. In addition, we will compare the proposed methods to several state-of-the-art deep learning-based methods.

1) IN Dataset: The IN dataset collected in northwestern India using the AVIRIS sensor in 1992, records remote sensing images of Indian Pines. The Indian Pines image is 145 × 145 in size, and includes 224 bands. Eliminating the noisy bands, only 200 hyperspectral bands are employed in the experiments. The ground truth of India Pines is divided into 16 categories, including Alfalfa, Corn, Woods, and so on, which are not all mutually exclusive.

2) PU Dataset: The PU dataset was collected using the ROSIS sensor at Pavia University in Italy. The spatial dimension of the Pavia University image is 610 × 340 with 103 spectral bands. This dataset contains nine categories, including Asphalt, Gravel, trees, and so on.

3) KSC Dataset: The AVIRIS sensor recorded the KSC located in the KSC area in Florida on March 23, 1996. The spatial dimensions of the KSC image are 512 × 614, and the spectral dimensions are 224 bands. After eliminating 48 noisy bands, there are 172 spectral bands left in the experiments. There are a total of 13 categories, including Scrub, Wate, Salt marsh, and so on.

4) Houston 2013 Dataset: The National Center for Airborne Laser Mapping on the University of Houston campus collected this hyperspectral data. After image processing, it was provided by the Geoscience and Remote Sensing Society data fusion competition in 2013. This image is 349 × 1905 in size, with 144 bands ranging from 364 to 1046 nm. There are a total of 15 categories, including Trees, Soil, Water, and so on.

B. Evaluation of Self-Supervised Features on the Target Classification Task

We conduct the following experiments to assess the performance of self-supervised pretext task methods on the target HSI classification task. We add a nonlinear classification layer, such as a softmax, at the bottom of CNN, and then train on the whole HSIs datasets for the IS pretext task method. Then, after initializing the parameters of CNN for the HSI classification task, we restore all pretrained weights and retrain the CNNs with five samples in each category by adding a new nonlinear classification layer. Finally, we evaluate the remained samples. The training samples are not augmented throughout the training process. It is noted that IS is short for a single pretext task method by predicting the IS and Spectral Order, and MT denotes the MT method.

1) Experiments on IN: Next, the experiment results of various pretext task methods on the Indian Pines testing dataset are reported. We compare the proposed methods to CNNs-based methods such as 2-D-CNN [15], 3-D-CNN [15], Hamida [16], and He [26], as well as the pretrained models (e.g., ResNet [28]), two contrastive instance learning methods (i.e., BYOL [56], SIMCLR [24]), and four commonly used SSL pretext task methods. Two series of evaluation metrics of all the methods on Indian Pines are presented in Table II. Table II shows the experimental results using five samples ($L = 5$) and ten samples ($L = 10$).

We can first observe that the proposed methods outperform other compared methods on both of the five and ten samples. It indicates the superiority of the proposed methods. A possible explanation for this might be that the CNN model learns a good feature representation using the proposed pretext task IS, resulting in a good performance on the target HSI classification task. Since the proposed multitask method contains many more labeled samples, it performs the best classification results. Second, the pretext task methods (i.e., IRO [44], IRE [50], IP [41], and DIM [60]) provide better performances than the CNNs-based methods, the pretrained method, and the contrastive instance learning methods. This result may be explained by the fact that the pretext task methods introduce more data prior to training the CNN model. Third, the contrastive instance learning methods and pretrained methods were observed to perform a bad performance when compared with the 2-D-CNN methods. The main reason is that the pretrained method (ResNet) is trained on the RGB images which are different to the HSIs. And the contrastive instance learning methods are designed to identify the RGB images and could not learn the subtle difference in the spectral dimension. Finally, the 2-D-CNN method outperforms the 3-D-CNNs based methods. It demonstrates that the superior of the 2-D-CNN method when there are only a few training samples. And these results also demonstrate the SSL methods could improve the performance of HSI classification.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II

CLASSIFICATION RESULTS OF IN

Class	2D-CNN		3D-CNN		Hamida		He		ResNet		BYOL		SIMCLR		IRO		IRE		IP		DIM		IS		MT	
	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10
1	94.70	92.10	45.90	70.60	45.80	51.50	40.00	51.50	23.08	62.63	26.37	22.14	29.75	45.26	70.18	90.00	93.50	98.60	85.70	96.00	72.00	95.90	59.50	89.86	88.89	83.78
2	45.90	60.30	35.10	41.80	22.60	32.60	43.10	33.60	49.83	50.07	48.29	54.60	45.28	42.51	43.29	65.90	44.10	59.70	46.60	58.10	43.80	59.00	54.82	70.53	73.30	77.86
3	38.50	60.70	43.20	43.80	20.70	22.40	20.40	30.00	35.30	39.85	51.32	48.77	31.68	44.56	26.98	62.10	38.30	54.20	49.10	61.80	44.20	61.00	55.81	64.83	52.62	63.62
4	59.10	61.30	35.30	37.70	21.70	32.30	27.00	47.00	53.90	48.50	51.39	56.60	45.48	53.87	45.42	61.80	46.70	59.80	46.30	54.40	58.00	65.40	44.86	68.85	73.36	71.39
5	57.40	66.00	38.90	47.50	31.10	47.60	38.60	49.60	55.26	73.60	48.48	58.94	55.26	58.45	55.91	66.80	37.10	64.30	46.70	61.50	55.40	69.50	64.90	82.03	65.26	65.47
6	86.60	96.40	61.30	79.60	72.30	85.90	74.40	83.40	73.97	80.66	87.70	88.76	78.16	90.52	92.17	95.10	88.80	94.70	85.20	93.80	83.80	94.70	86.39	94.08	81.40	92.09
7	45.00	37.10	32.40	37.50	21.30	24.20	25.60	32.70	14.57	100	30.51	53.06	16.67	39.39	33.33	47.40	56.20	61.00	34.60	50.70	48.00	53.70	35.29	41.27	65.45	65.00
8	94.70	94.50	82.30	85.60	87.20	92.70	79.90	92.20	83.05	91.38	87.17	84.77	83.84	84.85	91.23	94.70	92.90	94.70	93.70	94.50	93.00	94.70	91.04	94.20	94.51	92.87
9	10.80	23.50	21.40	19.80	36.10	27.40	8.00	14.80	24.10	17.86	12.05	11.49	11.30	17.24	18.07	28.60	8.70	19.80	11.60	18.90	11.30	22.20	18.87	6.17	14.39	16.67
10	44.80	58.00	45.50	56.30	44.70	44.10	44.20	53.00	39.48	45.88	46.77	47.34	49.25	41.55	59.63	63.50	56.90	69.90	51.40	63.70	48.40	61.60	66.70	68.68	59.98	74.59
11	58.20	75.50	45.00	65.40	57.00	58.70	43.80	60.90	58.18	62.79	53.19	64.23	64.95	61.69	60.81	76.80	67.40	76.40	65.70	78.30	65.60	78.50	74.52	72.09	76.90	79.98
12	41.40	54.60	35.10	42.10	21.10	35.90	29.60	33.90	37.89	47.45	47.39	46.08	38.74	55.30	38.88	57.10	39.30	55.00	38.00	48.00	41.50	52.10	22.90	55.53	59.98	69.78
13	82.80	90.20	59.00	74.50	57.40	78.30	77.00	75.00	49.81	87.94	88.58	85.58	70.07	90.03	86.34	98.20	70.30	92.80	80.90	89.80	81.20	95.10	64.93	87.76	92.82	88.32
14	89.20	89.50	83.40	85.60	77.60	81.40	84.10	86.60	82.25	89.29	87.46	87.26	87.14	93.70	88.79	91.80	85.30	89.90	89.30	90.90	87.90	91.50	87.49	92.73	89.80	92.01
15	16.80	42.10	0.00	42.80	0.00	26.90	15.90	25.50	40.12	56.71	49.73	40.12	50.00	51.51	8.15	52.60	14.80	49.80	20.80	42.90	16.10	49.10	47.77	54.78	51.23	56.66
16	77.00	75.10	51.90	67.90	50.00	73.80	39.20	59.60	56.45	73.93	78.05	80.41	56.34	75.62	45.60	72.20	65.60	60.40	69.20	75.10	81.40	70.10	79.23	93.59	6 6.94	82.80
OA (%)	56.53	67.79	47.60	57.52	45.61	51.18	46.99	53.71	52.56	60.30	56.87	60.55	55.98	60.14	57.86	70.40	57.91	68.78	58.83	68.13	58.34	69.35	64.36	71.44	**69.89**	**75.21**
κ (%)	51.40	64.00	42.40	52.80	39.20	45.70	41.30	48.50	47.29	55.47	52.28	56.00	50.81	55.06	52.67	66.90	52.90	65.10	53.90	64.40	53.20	65.70	60.17	68.16	**66.27**	**72.04**
Pre_training times (S)											288.84		498.96		588.68		416.96		363.57		396.62		2168.18		3601.33	
Parameters (MB)											14.88		0.94		0.49		0.49		0.49		1.66		0.49		0.49	

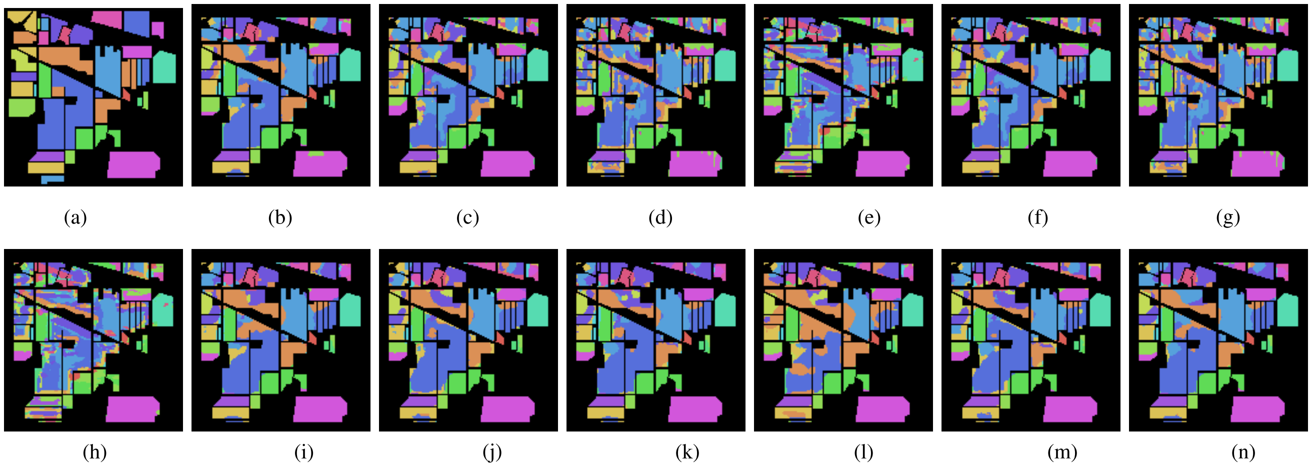

(a) (b) (c) (d) (e) (f) (g)
(h) (i) (j) (k) (l) (m) (n)

Fig. 10. Classification maps obtained on the AVIRIS Indian Pines dataset (with five training samples). (a) Ground truth. (b) 2-D-CNN. (c) 3-D-CNN. (d) Hamida. (e) He. (f) ResNet. (g) BYOL. (h) SIMCLR. (i) IRO. (j) IRE. (k) IP. (l) DIM. (m) IS. (n) MT.

We also performed another comparison experiment of all the SSL methods on the pre_training times and the parameters. We can see that five pretext task methods (e.g., IRO, IRE, IP, IS, and MT) have the same number of parameters, which are fewer parameters than other SSL methods. This result demonstrates that the pretext task method is easy to implement. However, the pretext task methods usually need much more time to learn a good representation. Among them, MT takes the most time to train the deep learning model. A possible explanation for this might be that MT generates much more samples for training the 2-D-CNN. For example, after data processing, IS generates 16 times data. And MT contains much more data from two pretext task methods (e.g., IS and IRO). Fig. 10 provides a visual comparison of the performances of all methods.

2) Experiments on Houston 2013 Dataset: We adjusted the number of parameters to 144 to match the bands of Houston data. We compare the proposed methods to the CNNs-based methods, pretrained ResNet method, contrastive instance learning methods (i.e., BYOL [56], SIMCLR [24]), and four frequently used pretext task SSL methods. We show the experiment results of all the methods in Table III. It is apparent that the proposed methods achieve the best performance, followed by the SSL methods (such as Image Rotation, IP, IRE and Image DIM methods), the 2-D-CNN method, the pretrained ResNet method, and the contrastive instance learning methods. This study confirms that our proposed methods can improve the performance of CNNs. The 3-D-CNNs-based methods perform the worst results, such as 3-D-CNN [15], Hamida [16], and He [26]. This is also because of the insufficient labeled training samples in training the 3-D-CNNs-based algorithms. Furthermore, it also indicates that training the 3-D-CNNs-based methods requires significantly more labeled data. We can easily observe that MT has fewer parameters, but needs much more time to learn a stable representation. The Houston classification results of all the methods are shown in Fig. 11.

3) Experiments on PU: In this experiment, we build the same CNNs structures as in the IN experiment, with the exception that the number of parameters is modified to match the 102 hyperspectral bands. In this experiment, we compare the proposed methods to the CNNs-based methods (i.e., deep recurrent neural network (DRNN) (1-D-CNN) [12], 2-D-CNN [15], Hamida [16], He [26]), one per-trained method

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

TABLE III

Classification Results of the Houston Data

Class	2D-CNN		3D-CNN		Hamida		He		ResNet		BYOL		SIMCLR [24]		IRO		IRE		IP		DIM		IS		MT	
	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10
1	87.52	89.44	80.04	91.31	65.66	0.00	55.89	61.69	80.09	83.51	48.27	61.57	49.75	61.39	83.52	94.42	88.71	87.46	85.70	92.01	94.18	84.13	89.73	93.01	90.48	91.54
2	82.44	82.85	68.80	90.76	50.97	38.20	31.30	26.94	64.04	79.89	26.03	51.69	27.39	47.97	84.96	91.71	84.59	89.39	71.70	86.48	82.77	80.34	86.22	94.36	84.39	92.77
3	97.57	98.02	88.16	91.66	64.38	76.99	38.77	30.92	68.62	79.91	45.20	90.36	58.94	70.56	98.40	97.70	97.64	99.85	96.30	99.93	94.85	97.37	96.24	99.56	99.41	99.34
4	93.23	92.05	62.76	93.27	53.52	3.35	40.72	14.30	74.96	81.88	67.62	79.59	73.73	69.16	81.46	94.83	83.93	90.91	75.54	88.95	89.45	92.98	94.80	96.69	91.89	96.21
5	90.08	95.00	81.22	87.43	68.94	51.83	68.63	33.03	78.78	89.44	66.96	69.39	63.26	80.91	91.01	96.45	87.36	95.96	87.15	95.54	92.25	96.03	95.54	91.28	97.43	98.15
6	80.15	92.12	74.54	87.04	19.13	40.81	26.91	19.45	61.98	80.84	37.91	54.07	41.38	51.09	84.90	91.58	59.74	92.92	89.19	91.57	76.14	86.80	91.44	88.67	85.25	94.77
7	74.52	88.48	78.25	85.35	40.41	53.08	38.69	33.95	69.11	80.98	74.44	84.01	69.96	78.37	68.76	87.76	78.02	85.26	81.17	89.18	82.35	85.31	85.14	89.88	88.59	93.24
8	57.22	49.70	64.25	55.94	31.96	58.98	37.98	7.67	39.23	57.16	55.02	54.95	54.53	50.81	49.56	68.10	38.48	72.39	41.00	66.67	36.00	72.58	54.96	72.56	76.47	81.39
9	75.71	80.41	80.55	70.88	35.00	42.59	25.52	22.19	43.20	69.05	67.84	77.13	60.60	69.35	64.93	81.83	82.74	81.25	72.01	79.24	77.17	80.95	81.40	86.43	76.16	81.55
10	57.64	75.15	53.06	49.21	31.19	33.94	16.90	25.30	32.68	63.07	50.45	55.79	37.97	42.79	67.84	70.28	52.92	73.39	42.98	64.28	55.29	81.10	61.75	86.49	80.25	85.89
11	59.07	79.66	40.63	72.43	21.91	3.53	25.27	37.81	43.28	65.49	72.63	85.93	61.08	70.91	66.20	81.08	38.81	75.16	51.35	57.00	63.44	71.77	80.04	86.88	79.11	89.87
12	44.15	78.23	28.01	56.67	46.49	4.45	16.90	26.04	55.47	64.31	57.04	59.49	57.85	73.84	48.31	68.11	62.84	80.52	68.73	65.58	58.22	65.56	58.48	76.77	79.58	84.38
13	88.94	87.45	83.08	86.90	35.79	20.66	23.16	52.20	51.76	74.77	86.25	83.06	77.83	86.85	63.88	91.87	81.54	89.82	82.92	90.46	85.33	93.51	84.39	94.17	86.08	94.53
14	84.54	99.41	85.21	92.99	59.80	46.15	28.33	26.35	55.26	83.53	89.31	89.18	85.11	87.96	100	99.76	96.45	93.72	92.74	99.40	95.56	97.57	86.16	98.54	94.32	88.86
15	92.65	96.15	67.59	84.21	64.47	80.50	74.16	32.63	60.27	75.76	66.11	80.27	42.46	72.25	88.54	96.65	97.10	94.99	84.83	97.74	88.50	97.44	95.02	96.27	96.15	97.14
OA (%)	74.76	83.86	67.77	77.54	47.73	37.34	39.19	30.80	58.96	74.37	59.45	70.20	56.97	66.32	74.11	85.11	73.75	84.89	72.10	81.32	75.59	83.51	80.62	88.67	**85.62**	**90.22**
κ (%)	72.73	82.57	65.22	75.75	43.60	32.12	34.56	25.79	55.75	72.32	56.28	67.83	53.57	63.70	72.06	83.92	71.66	83.68	69.88	79.81	73.63	82.19	79.07	87.76	**84.46**	**89.43**
Pre_training times (S)											51.44		87.13		871.53		655.14		612.54		611.29		290.66		477.81	
Parameters (MB)											14.88		0.94		0.45		0.45		0.45		1.63		0.45		0.45	

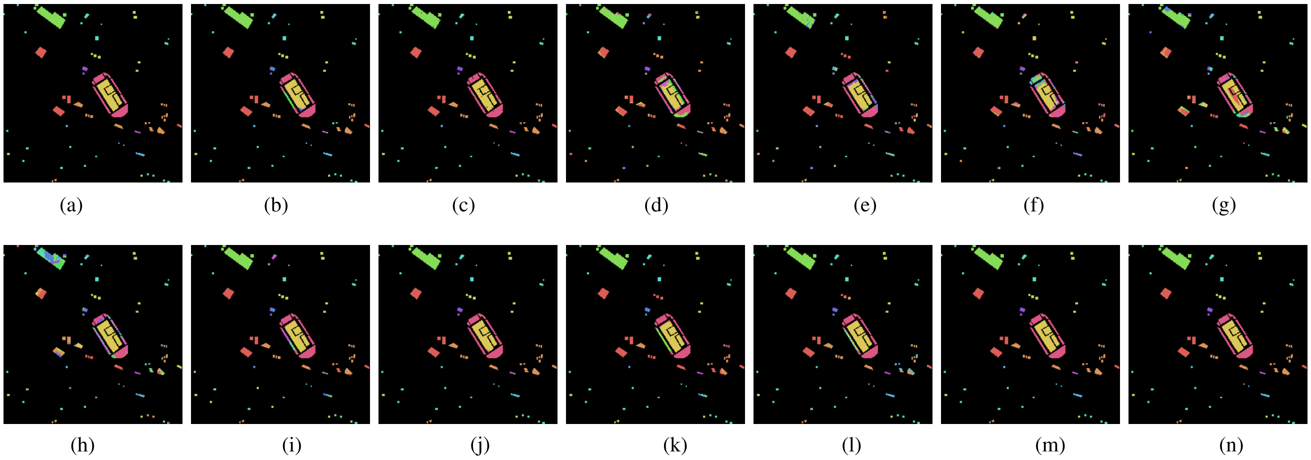

(a) (b) (c) (d) (e) (f) (g)
(h) (i) (j) (k) (l) (m) (n)

Fig. 11. Classification maps obtained on some areas of the Houston dataset (with ten training samples). (a) Ground truth. (b) 2-D-CNN. (c) 3-D-CNN. (d) Hamida. (e) He. (f) ResNet. (g) BYOL. (h) SIMCLR. (i) IRO. (j) IRE. (k) IP. (l) DIM. (m) IS. (n) MT.

TABLE IV

Classification Results of the PU

Class	DRNN		2D-CNN		Hamida		He		ResNet		BYOL		SIMCLR		IRO		IRE		IP		DIM		IS		MT	
	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10
1	81.10	79.70	57.40	74.80	47.00	63.00	14.30	43.50	52.74	67.68	74.08	71.30	68.49	75.63	70.00	75.30	55.30	72.30	58.70	73.90	52.60	74.70	83.45	78.63	82.75	84.20
2	53.19	77.70	70.50	79.30	78.40	79.80	77.60	80.50	79.59	79.65	65.52	74.50	75.02	81.06	67.40	83.20	72.20	80.60	72.10	80.40	71.70	81.50	75.56	87.20	83.74	89.34
3	44.26	67.10	53.50	63.00	29.70	42.00	16.10	26.10	45.03	36.10	53.57	49.80	45.61	54.97	65.70	57.80	56.80	58.20	54.90	62.40	53.30	54.10	64.52	62.48	75.99	74.89
4	64.20	69.60	92.50	95.10	73.40	76.80	59.90	83.60	71.25	77.75	59.93	54.64	49.63	60.41	89.20	95.40	93.70	94.70	94.10	95.30	92.80	94.50	94.92	95.47	92.91	92.66
5	98.67	94.70	99.10	98.70	93.70	97.00	87.00	61.20	93.26	81.84	99.22	98.83	96.79	99.59	99.70	99.00	99.70	99.60	99.70	99.00	99.10	99.10	98.95	99.74	98.91	99.66
6	32.17	54.80	52.00	68.90	41.10	22.60	23.20	55.50	48.87	38.26	46.22	31.07	49.21	50.09	54.40	76.50	56.90	71.50	56.10	69.90	53.20	71.60	64.33	78.36	61.61	86.42
7	40.56	56.70	46.00	67.10	37.80	56.50	19.90	34.00	47.17	52.98	45.76	48.58	51.09	58.83	63.30	67.90	49.80	73.30	47.80	64.10	43.20	71.90	62.82	80.57	80.41	75.62
8	41.85	69.90	68.00	73.10	49.50	62.20	31.50	54.60	62.20	59.07	76.06	74.97	72.81	75.60	73.00	73.70	69.50	69.80	67.00	75.50	71.50	69.80	78.48	81.01	84.91	86.04
9	96.72	97.00	83.30	92.40	78.80	71.50	53.60	91.90	82.27	68.58	70.00	65.95	53.53	72.10	82.90	94.10	96.90	96.40	90.10	95.30	86.40	94.30	96.69	96.99	90.77	97.90
OA (%)	55.52	72.69	63.16	73.98	58.98	65.39	50.44	62.37	64.85	65.14	61.44	63.41	64.35	69.75	65.80	76.45	65.28	74.21	64.88	74.49	63.38	74.42	73.53	80.75	**79.02**	**84.10**
κ (%)	46.46	65.90	55.80	68.00	48.40	54.70	35.40	52.00	54.98	54.72	53.08	54.00	54.15	61.75	59.20	70.80	58.30	68.30	57.90	68.60	56.10	68.40	67.54	75.40	**72.91**	**79.63**
Pre_training times (S)											1127.14		2059.82		1787.58		1291.44		1187.45		1482.68		6134.57		9683.75	
Parameters (MB)											14.88		0.94		0.43		0.43		0.43		1.61		0.43		0.43	

(such as ResNet [28]), two contrastive instance learning methods (i.e., BYOL [56], SIMCLR [24]), and four commonly pretext task methods. Table IV displays the experiment results of all methods. From Table IV, we can observe again that the proposed methods outperform other comparison methods both on five and ten training samples. It demonstrates the superiority of the proposed methods. In addition, the self-supervised pretext task methods outperform all other methods. A possible explanation for these results may be the self-supervised pretext task methods could offer much more prior knowledge and learn a good feature representation. Since the difference between HSIs and RGB images, the pretrained method (such as ResNet) and the contrastive instance learning methods fail to perform a better performance. The 3-D-CNNs based methods again perform the worst classification results. It indicates that the CNNs-based methods urgently need much more training samples. Again, we also observe that MT and IS have fewer parameters, but take a lot of time to learn the representation. Finally, the visualizations of all the approaches are shown in Fig. 12.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

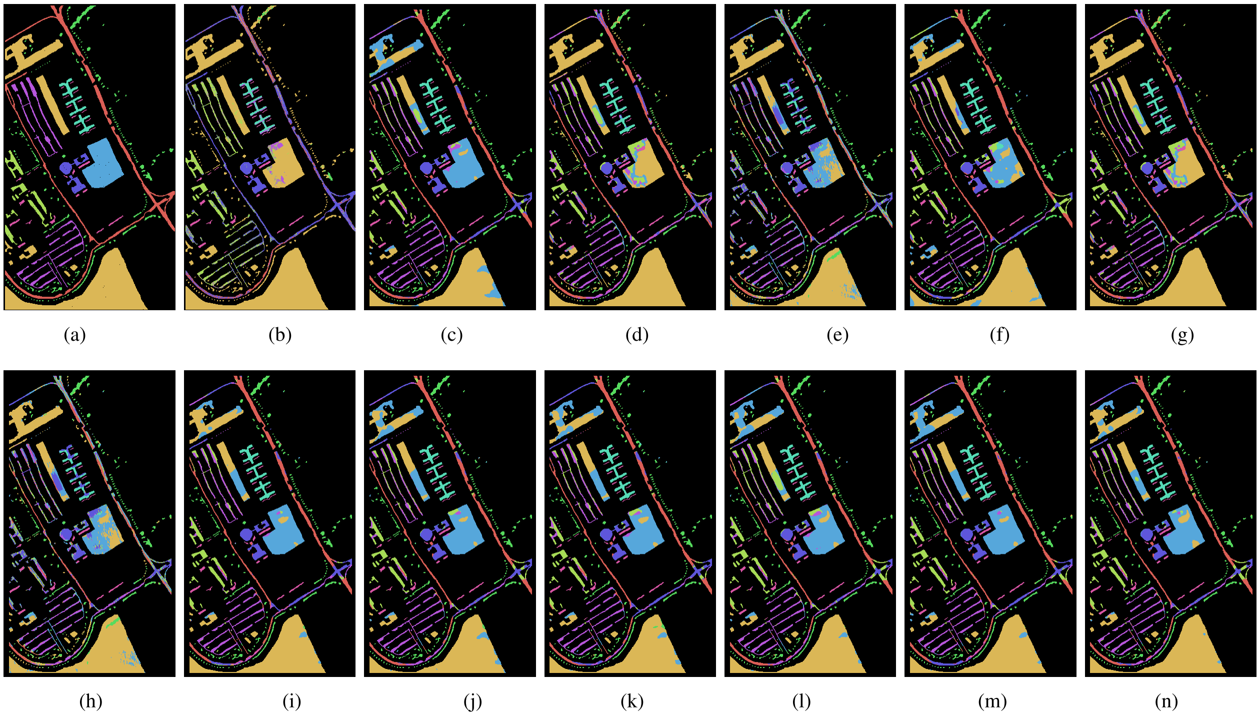


Fig. 12. Classification maps obtained on the PU dataset (with five training samples). (a) Ground truth. (b) DRNN. (c) 2-D-CNN. (d) Hamida. (e) He. (f) ResNet. (g) BYOL. (h) SIMCLR. (i) IRO. (j) IRE. (k) IP. (l) DIM. (m) IS. (n) MT.

TABLE V

CLASSIFICATION RESULTS OF THE KSC

Class	2D-CNN		3D-CNN		Hamida		He		ResNet		BYOL		SIMCLR		IRO		IRE		IP		DIM		IS		MT	
	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10	L=5	L=10
1	8.21	83.00	82.00	83.20	0.00	0.00	34.00	25.70	75.08	73.90	81.73	85.37	74.20	83.08	80.00	83.70	81.00	83.10	81.00	83.60	82.00	82.70	85.43	85.32	84.19	84.70
2	61.00	60.50	53.00	69.10	44.00	43.90	16.00	38.60	24.14	30.26	80.21	77.89	65.04	87.77	64.00	82.10	64.00	76.80	60.00	71.50	52.00	75.00	52.63	87.01	76.89	83.78
3	40.39	49.40	71.00	88.60	6.00	20.90	61.00	61.90	40.81	56.57	90.47	91.81	87.68	90.69	80.00	92.00	70.00	93.20	68.00	85.80	70.00	89.30	93.13	83.94	86.44	92.50
4	29.48	0.80	47.00	62.40	10.00	22.20	25.00	41.80	26.92	43.58	52.65	65.89	45.63	75.78	51.00	72.00	50.00	68.00	50.00	62.00	51.00	64.50	64.23	67.58	55.61	77.21
5	36.41	45.20	66.00	86.70	51.00	56.20	46.00	58.30	45.70	54.04	83.60	89.44	59.26	89.68	68.00	86.30	71.00	90.70	70.00	87.60	71.00	83.70	65.69	86.67	87.65	84.44
6	28.63	10.00	88.00	92.50	17.00	25.70	43.00	42.30	60.00	69.59	74.53	89.05	65.04	80.21	87.00	92.20	87.00	92.80	90.00	92.70	87.00	95.20	95.63	86.29	84.14	94.96
7	43.24	41.80	95.00	99.50	23.00	35.20	47.00	54.30	77.60	93.62	100	96.26	84.06	99.42	83.00	97.90	82.00	95.50	84.00	99.50	90.00	100	80.51	100	98.45	97.21
8	32.93	21.70	62.00	89.50	7.00	11.70	56.00	67.60	58.38	68.44	54.56	72.24	57.18	81.92	70.00	92.10	68.00	86.50	66.00	91.60	67.00	90.00	68.19	87.56	81.39	86.54
9	64.41	68.30	88.00	97.90	42.00	33.50	80.00	88.30	69.89	86.81	77.38	87.55	76.92	96.46	87.00	98.10	89.00	94.80	88.00	97.80	89.00	97.70	87.75	94.90	90.65	99.60
10	51.24	10.40	48.00	83.70	33.00	51.90	22.00	44.20	45.06	58.01	44.76	87.48	84.09	74.02	52.00	86.60	52.00	86.40	51.00	89.50	47.00	85.80	76.96	84.04	94.61	91.35
11	91.90	91.80	100	99.60	25.00	33.10	70.00	89.30	97.14	92.43	96.86	97.09	93.98	97.56	99.00	99.80	99.00	99.80	100	99.50	99.00	99.00	97.65	99.25	96.98	98.62
12	20.28	17.30	66.00	80.90	46.00	49.90	46.00	49.30	41.84	44.58	78.84	94.97	91.02	88.08	64.00	85.60	69.00	83.80	66.00	85.50	69.00	83.10	72.39	90.87	88.42	88.54
13	77.24	77.40	97.00	97.90	59.00	62.40	75.00	88.60	93.44	94.61	97.35	100	99.78	99.94	95.00	97.70	95.00	97.60	96.00	99.00	97.00	98.10	95.36	99.07	97.68	99.45
OA (%)	48.15	58.16	75.96	86.68	33.21	36.59	53.96	61.52	64.55	69.75	77.11	86.98	78.57	87.36	77.24	88.82	77.15	87.42	76.60	87.94	76.78	87.46	81.15	88.24	**86.60**	**89.93**
κ (%)	42.50	53.00	73.00	85.20	27.00	30.00	50.00	57.80	60.82	66.57	74.66	85.58	76.34	85.97	75.00	87.60	75.00	86.10	74.00	86.60	74.00	86.10	79.17	86.96	**85.17**	**88.84**
Pre_training times (S)											189.69		279.91		495.64		242.49		214.82		255.81		1265.94		2130.46	
Parameters (MB)											14.88		0.94		0.47		0.47		0.47		1.65		0.47		0.47	

4) Experiments on KSC : We set the number of parameters corresponding to the 176 spectral bands, and the experimental results of all methods are reported on Table V. We compare the proposed methods to CNNs-based methods (i.e., 2-D-CNN [15], 3-D-CNN [15], Hamida [16], and He [26]), pretrained methods (such as ResNet [28]), two contrastive instance learning methods (i.e., BYOL [56], SIMCLR [24]), as well as four frequently used pretext task SSL methods. Table V shows the results of all methods on the KSC dataset. It is clear that the proposed pretext task methods again outperform the other comparison methods. The SSL methods (i.e., Image Rotation [44], IP [41], IRE [50], and Image DIM [60] methods) are observed to produce better performances than other comparison methods. The 3-D-CNNs-based methods, produce the worst results, such as 3-D-CNN [15], Hamida [16], and He [26]. This is mainly because the labeled training samples are not enough to train the 3-D-CNNs-based methods. From Table III, we could also see that the proposed two methods take a lot of time to train the 2-D-CNN model, while having fewer parameters. The KSC classification results of all the methods are shown in Fig. 13.

C. Ablation Study of SSL Tasks

In this section, we conduct several experiments on five training samples to learn the performance of target HSI classification with various SSL pretext task methods, including the choice of single self-supervised pretext task method and the domain difference between source and target dataset. It should be noted that IN and PU stand for Indian Pines and PaviaU, respectively.

1) Ablation Study of Nonoverlap Measurement: In addition to the SSL stage, the overlap of samples also contributes to the

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

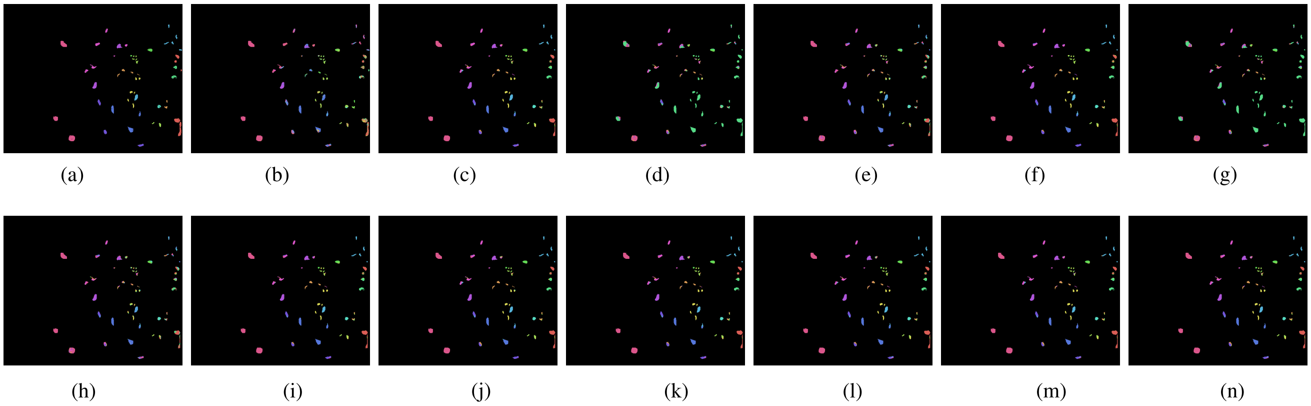


Fig. 13. Classification maps obtained on the KSC dataset (with five training samples). (a) Ground truth. (b) 2-D-CNN. (c) 3-D-CNN. (d) Hamida. (e) He. (f) ResNet. (g) BYOL. (h) SIMCLR. (i) IRO. (j) IRE. (k) IP. (l) DIM. (m) IS. (n) MT.

TABLE VI

Classification Results on All Testing Samples (Overlap) and Nonoverlapped Testing Samples (Nonoverlap)

Methods	Non-overlap		Overlap	
	2D-CNN	MT	2D-CNN	MT
IN	52.65	**64.19**	57.52	**75.21**
PU	71.06	**81.30**	77.54	**90.22**
KSC	75.32	**80.01**	86.68	**89.93**
Houston	63.98	**81.39**	73.98	**84.10**

TABLE VII

Results of Different Methods

Methods	OA on the target classification task			
	IN	PU	KSC	Houston
2D-CNN	54.46 ± 3.44	61.80 ± 5.29	72.26 ± 2.74	74.67 ± 3.50
P-2D-CNN	55.15 ± 3.38	65.39 ± 6.20	74.73 ± 2.22	76.14 ± 2.58
IS	**56.81 ± 3.97**	**67.41 ± 5.85**	**76.40 ± 2.17**	**79.73 ± 2.03**

improvement in classification accuracy. To demonstrate how overlapping samples affect the efficacy of SSL, we conducted the experiment with nonoverlapping testing and training samples.

Using the sampling strategy for evaluation [61], we remove testing samples that overlapped with training samples. Then we utilize only nonoverlapping testing samples with the same parameters as previous experiments. Table VI details the classification findings of all testing samples (Overlap) and those that do not overlap (Nonoverlap). The 2-D-CNN is employed as the baseline method. The findings of the experiment indicate that whether testing on nonoverlap or overlap samples, performance is improved when the SSL step is used. This outcome also illustrates the efficacy of the SSL phase. In the four hyperspectral datasets, we can also notice that the classification accuracy of Overlap is superior to that of Nonoverlap. This result is primarily attributable to the spatial information overlap between the training and testing samples. Consequently, it is necessary to further investigate ways for generating more reliable classification findings. In contrast, the proposed MT categorization actuary works surprisingly well with the nonoverlap. On the basis of nonoverlapping testing samples, we will investigate the potential for the SSL to improve classification accuracy.

2) Ablation Study of the Proposed IS Method: We conduct the experiments and evaluate the performance of different methods, i.e., 2-D-CNN, Pretrained 2-D-CNN (P-2-D-CNN), IS. Table VII presents the average results of the experiments that are retrained on four benchmark HSIs datasets ten times. It is noted that we adopt the Houston dataset to P-2-D-CNN for Indian Pines, PaviaU, and KSC datasets, but the PaviaU dataset to pretrain the Houston dataset. What stands out in Table VII is that the proposed IS method outperforms the other methods. This demonstrates that the proposed IS method is an effective and simple way to improve performance. Not surprisingly, the P-2-D-CNN produces a better performance than 2-D-CNN, which causes by introducing the external labeled data. Since the big difference between different HSIs datasets, the P-2-D-CNN could not outperform IS. The 2-D-CNN provides the worst classification results because of the insufficient samples. These observational studies suggest that the SSL method in HSIs may help small dataset classification and improve performance.

3) Ablation Study of the Choice of SSL Pretext Task Method: We conduct the experiments and evaluate the feature learning performance of several self-supervised pretext task methods on target HSIs categorization. The experimental results on four benchmark HSIs datasets are shown in Table VIII. The pretrained parameters acquired from pretext task methods to initialize the target HSI classification task. And the model is retrained for ten times. Table VIII shows that IS pretext task method outperforms the other four self-supervised pretext task methods. This demonstrates IS pretext task method learns a more suitable feature representation using the difference between the central pixel and all boundary pixels. These results indicate that it is critical to learn the high-level feature representations for HSI classification by selecting an appropriate pretext task method. And, IS pretext task method can provide a good feature representation for the HSI classification.

4) Ablation Study of Different Max Scale Size: Apart from the learnable parameters of the CNN and the hyperparameters, the max scale size of the proposed method plays a significant role in the classification performance. Therefore, we conduct the experiments with the different scales

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
RESULTS OF THE CHOICE OF SSL PRETEXT TASK METHODS

Pretext task method	OA on the target classification task			
	IN	PU	KSC	Houston
IRO [44]	53.59 ± 3.37	63.01 ± 5.46	75.10 ± 4.30	76.94 ± 1.43
IRE [50]	55.80 ± 6.21	62.19 ± 6.13	76.19± 3.57	72.62 ± 1.47
IP [41]	52.65 ± 3.50	65.87 ± 4.75	74.32 ± 6.01	68.66 ± 7.36
DIM [60]	53.14 ± 3.31	75.79 ± 4.90	73.78 ± 0.70	65.52 ± 6.26
IS	**56.81 ± 3.97**	**61.62 ± 3.82**	**78.11 ± 1.64**	**81.28 ± 1.99**

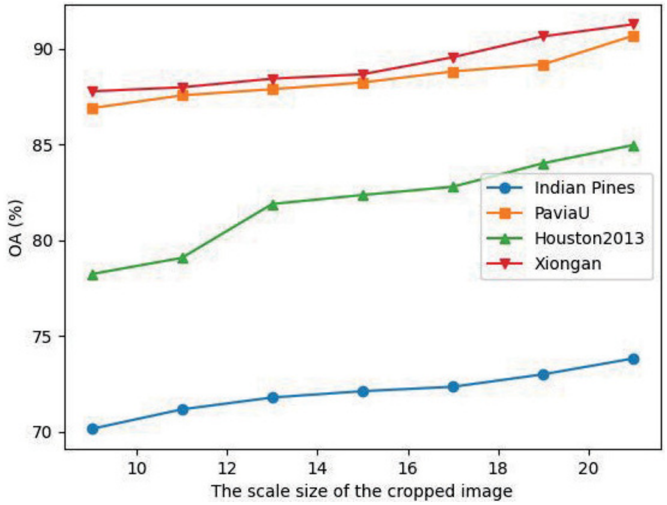
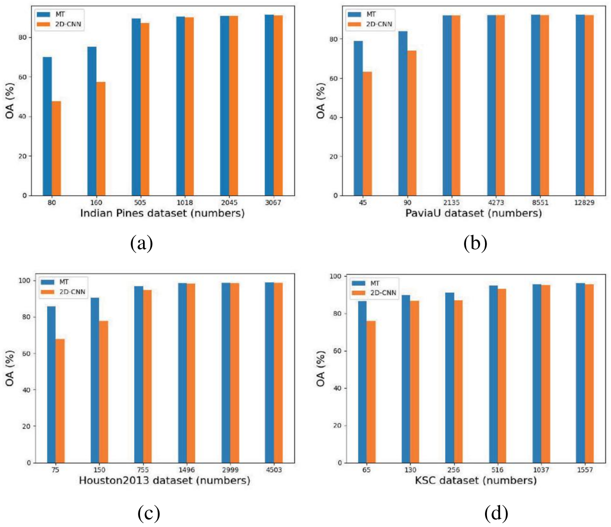


Fig. 14. Results of different max scale sizes.

(i.e., 9, 11, 13, 17, 19, 21) to evaluate the performance of the proposed method. The results of OA obtained by IS are reported in Fig. 14. It is noted that the generated images should be resized to the max size before feeding into the CNN. We can observe that the performance of the CNN increases with varying the max scale size. A possible explanation for this might be that with more max scale size, there are much more training samples that could lead to learning a good feature representation for the target classification task.

5) Ablation Study of Different Number of Training Samples: The SSL method could not only improve the performance of HSI classification, but also relieve the burden of training on a large number of supervised labeled data. In this section, we conduct experiments with the different training samples to evaluate the effectiveness of the SSL method. During the experiments, we first pick out 5 and 10 from each category of different HSIs datasets, resulting in 45, 65, 75, and 80 training samples of PaviaU, KSC, Houston2013, and Indian Pines datasets. We then vary samples from 0.05% to 0.3% to obtain different number of training samples of different HSIs datasets (as shown in Fig. 15). Using these different number of training samples, the results of OA obtained by MT and 2-D-CNN are shown in Fig. 15. We can see that the performances both of MT and 2-D-CNN are improved with increasing the number of training samples. Moreover, MT (SSL method) outperforms 2-D-CNN (supervised method). Finally, the 2-D-CNN performs satisfactory classification results with many more training samples. However, MT achieves comparable classification results to the 2-D-CNN with fewer samples. It demonstrates the SSL step could relieve the burden of training on a large number of supervised labeled data.



(a)

(b)

(c)

(d)

Fig. 15. Classification results of different number of training samples. (a) Indian Pines dataset. (b) PaviaU dataset. (c) Houston2013 dataset. (d) KSC dataset.

V. CONCLUSION

In this article, we first introduced self-supervised pretext task approaches for HSI classification by utilizing limited labeled examples. Then, for HSI classification, we suggested a simple SSL method called IS to predict the IS and Spectral Order. On the other hand, we developed an MT method by integrating two pretext task methods (such as IRO and IS) and the contrastive loss function. Additionally, using four benchmark HSIs datasets, we conducted a fair comparison of the suggested approaches with the pretrained methods, the contrastive instance learning methods, the supervised CNNs-based methods, and four well-utilized signal pretext task methods. According to the experimental results, SSL pretext task approaches outperform other comparison methods that use limited labeled samples. Additionally, we conducted ablation studies to analyze the effect of IS and MT methods on HSI classification, and discovered that the choice of self-supervised pretext tasks and the domain difference between the source and target datasets affect the performance of IS. We also prove that IS and MT could relieve the burden of training on a large number of supervised labeled data. Future work will continue to examine SSL and expand its application by developing a robust SSL framework for a variety of HSIs sectors to promote the adoption of SSL methods by HSIs organizations.

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[2] A. J. Brown, "Spectral curve fitting for automatic hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1601–1608, Jun. 2006.

[3] V. E. Brando and A. G. Dekker, "Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1378–1387, Jun. 2003.

[4] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

[5] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Appl. Sci.*, vol. 9, no. 10, p. 2110, 2019.

[6] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[7] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[8] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[9] C. Tao, W. Lu, J. Qi, and H. Wang, "Spatial information considered network for scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 984–988, Jun. 2021.

[10] X. Yang et al., "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 2033, 2020.

[11] 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.

[12] B. Liu, X. Yu, P. Zhang, X. Tan, A. Yu, and Z. Xue, "A semi-supervised convolutional neural network for hyperspectral image classification," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 839–848, Sep. 2017.

[13] Y. Zhang, Y. Wang, X. Chen, X. Jiang, and Y. Zhou, "Spectral-spatial feature extraction with dual graph autoencoder for hyperspectral image clustering," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 5, 2022, doi: 10.1109/TCSVT.2022.3196679.

[14] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[15] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[16] 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.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[19] O. A. B. Penatti, K. Nogueira, and J. A. D. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.

[20] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[21] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.

[22] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3318–3329, Jul. 2020.

[23] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[24] 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.

[25] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[26] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3904–3908.

[27] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] L. Ran, Y. Zhang, W. Wei, and T. Yang, "Bands sensitive convolutional network for hyperspectral image classification," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Aug. 2016, pp. 268–272.

[30] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, Feb. 2017.

[31] V. Sharma, A. Diba, T. Tuytelaars, and L. Van Gool, "Hyperspectral cnn for image classification & band selection, with application to face recognition," KU Leuven, ESAT, Leuven, Belgium, Tech. Rep. KUL/ESAT/PSI/1604, 2016.

[32] P. R. Lorenzo, L. Tulczyjew, M. Marcinkiewicz, and J. Nalepa, "Hyperspectral band selection using attention-based convolutional neural networks," *IEEE Access*, vol. 8, pp. 42384–42403, 2020.

[33] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, 2017.

[34] H. Lee and H. Kwon, "Contextual deep CNN based hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3322–3325.

[35] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Jul. 2016.

[36] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4L: Self-supervised semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1476–1485.

[37] I. Misra and L. Van Der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6707–6717.

[38] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 22, 2021, doi: 10.1109/TKDE.2021.3090866.

[39] P. Liu, M. Lyu, I. King, and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4571–4580.

[40] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[41] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[42] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," 2014, *arXiv:1410.8516*.

[43] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 577–593.

[44] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–16.

[45] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10334–10343.

[46] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 567–575.

[47] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2229–2238.

[48] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6391–6400.

[49] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9339–9348.

[50] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.

[51] Y. Wang et al., "Self-supervised feature learning with CRF embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2628–2642, May 2019.

[52] 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. (CVPR)*, Jun. 2020, pp. 9729–9738.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.

[53] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[54] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.

[55] 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.*, vol. 33, 2020, pp. 9912–9924.

[56] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

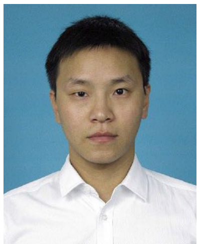
[57] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[58] 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, pp. 1–13, 2022.

[59] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 776–794.

[60] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–24.

[61] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, Feb. 2017.



Xiaofei Yang received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2014 and 2019, respectively.

He currently holds a post-doctoral position with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include semisupervised learning, deep learning, remote sensing, transfer learning, and graph mining.

Weijia Cao received the master's and Ph.D. degrees in computer science from the University of Macau, Macau, China, in 2013 and 2017, respectively.

She is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. She is a Research Fellow at the Department of Computer and Information Science, University of Macau, Macau, China, and also with Yangtze Three Gorges Technology and Economy Development Company Ltd., Beijing. Her main research interests include multimedia encryption, machine learning, remote sensing, and image processing.

Yao Lu received the B.S. degree in computer science and technology from Huaqiao University, Xiamen, China, in 2015, and the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Harbin, China, in 2020.

She was a Post-Doctoral Fellow with the University of Macau, Macau, China, from 2020 to 2021. She is currently an Assistant Professor with the Biocomputing Research Center, Harbin Institute of Technology, Shenzhen, China. Her research interests include pattern recognition, deep learning, computer vision, and relevant applications.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized one of "Highly Cited Researchers" in 2020 and 2021. He serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEUTRAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Authorized licensed use limited to: Universidade de Macau. Downloaded on December 27,2022 at 08:34:20 UTC from IEEE Xplore. Restrictions apply.