A 3-D-Swin Transformer-Based Hierarchical Contrastive Learning Method for Hyperspectral Image Classification

Xin Huang, *Senior Member, IEEE*, Mengjie Dong, Jiayi Li, *Senior Member, IEEE*, and Xian Guo

***Abstract*— Deep convolutional neural networks have been dominating in the field of hyperspectral image (HSI) classification. However, single convolutional kernel can limit the receptive field and fail to capture the sequential properties of data. The self-attention-based Transformer can build global sequence information, among which the Swin Transformer (SwinT) integrates sequence modeling capability and prior information of the visual signals (e.g., locality and translation invariance). Based on SwinT, we propose a 3-D SwinT (3DSwinT) to accommodate the 3-D properties of HSI and capture the rich spatial–spectral information of HSI. Currently, supervised learning is still the most commonly used method for remote sensing image interpretation. However, pixel-by-pixel HSI classification demands a large number of high-quality labeled samples that are time-consuming and costly to collect. As unsupervised learning, self-supervised learning (SSL), especially contrastive learning, can learn semantic representations from unlabeled data and, hence, is becoming a potential alternative to supervised learning. On the other hand, current contrastive learning methods are all single level or single scale, which do not consider complex and variable multiscale features of objects. Therefore, this article proposes a novel 3DSwinT-based hierarchical contrastive learning (3DSwinT-HCL) method, which can fully exploit multiscale semantic representations of images. Besides, we propose a multiscale local contrastive learning (MS-LCL) module to mine the pixel-level representations in order to adapt to downstream dense prediction tasks. A series of experiments verify the great potential and superiority of 3DSwinT-HCL.**

***Index Terms*— Contrastive learning, hyperspectral image (HSI) classification, self-supervised learning (SSL), Swin Transformer (SwinT), Transformer.**

I. Introduction

HYPERSPECTRAL imaging, which combines imaging and spectroscopic techniques to detect spatial and spectral information of ground targets, is one of the most important remote sensing (RS) imaging techniques. A hyperspectral image (HSI) has narrow, continuous spectral bands, and a broad electromagnetic spectrum, allowing object identification and detection at fine-grained scales. HSI classification aims to assign predefined labels to each pixel and has been successfully applied to ecological science, urban planning, precision agriculture, and mineral exploration [1], [4].

Early hyperspectral classifications mostly focused on spectral feature extraction algorithms, such as support vector machines (SVMs) [5], random forests (RFs) [6], and logistic regression [7]. Considering the issue of dimensionality curse due to the inherent high-dimensional spectral bands of HSI, researchers investigated dimensionality reduction techniques. The principal component analysis (PCA) [8] attempted to find the optimal transformation to project the high-dimensional data into a low-dimensional subspace, which belongs to feature extraction. On the other hand, feature selection (or band selection) aimed to select the representative band subset from the original data [9], [10], which can effectively alleviate the computational burden and improve the efficiency of HSI classification. Unsupervised feature selection was more widely used considering the difficulty of sample collection [11]. However, generally, the spectral-classification methods did not achieve satisfactory results owing to the lack of spatial and contextual information.

Spatial features were capable of improving the representation capability of the hyperspectral data and enhancing the robustness of the model [2]. For example, in [12], the composite kernels could take into account both spectral and spatial signatures and, at the same time, balance their information, which improved the classification performance. The edge-preserving filtering method proposed by [13] was able to smooth the probability maps of SVM classifications in a postprocessing manner. The support tensor machine (STM) [14] aimed to characterize the information of classes in the tensor space of HSI, which can preserve the original spectral–spatial structures and alleviate the spatial information loss, compared to SVM. In addition, the superpixel segmentation technique [15] attempted to divide HSIs into local homogeneous regions of different sizes, enhance the consistency of spatial structure information, and eliminate the effect of salt-and-pepper noise. Existing studies showed that the superpixel segmentation technique can effectively extract the spatial information

Manuscript received 13 June 2022; revised 28 July 2022; accepted 23 August 2022. Date of publication 26 August 2022; date of current version 13 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41971295 and Grant 42071311, and in part by the Special Fund of the Hubei Luojia Laboratory under Grant 220100031. *(Corresponding author: Jiayi Li.)*

Xin Huang is with the School of Remote Sensing and Information Engineering and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn).

Mengjie Dong and Jiayi Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: mjdong@whu.edu.cn; zjjercia@whu.edu.cn).

Xian Guo is with the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China (e-mail: guoxian@bucea.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3202036

of HSI. Zhang *et al.* [10] adopted superpixel segmentation to construct different homogeneous regions and represented the correlations between neighboring pixels of homogeneous regions in the structure graph to preserve the spatial structure information. Inspired by Jiang *et al.* [15], Zhang *et al.* [16] proposed a novel spectral–spatial and SuperPCA method. The superpixel-based local reconstruction technique can make full use of spatial information and extract global–local contextual features simultaneously. In addition, it was found that the combination of the superpixel technique and sparse representation [17] could effectively utilize spectral and spatial information for HSI classification [18].

The above methods rely heavily on the domain knowledge and experience of human experts and suffer from the disadvantages of low generalization ability and limited characterization capability. Deep spectral–spatial feature extraction methods based on deep learning have become increasingly popular in RS over the last few years. Many researchers have attempted to transfer the powerful feature extraction capability of neural networks to HSI classification [19], [20], [21], [22], [23] and obtained promising results. As opposed to traditional methods, deep learning avoids the design of artificial features and can adaptively extract abstract high-level features from original data.

Among existing deep learning models, convolutional neural networks (CNNs) have received wide attention because of their strengths in weight sharing and local connectivity, which greatly reduces model complexity and reliance on spatial relationships. Given that HSI has distinct properties from natural images, researchers are devoted to constructing specialized CNN-based feature extractors to mine rich spatial–spectral information from HSI. For example, Slavkovikj *et al.* [24] proposed a CNN-based feature learning framework to extract structured information from HSI. Li *et al.* [25] developed a novel pixel-pair approach to enhance the recognition capability of CNN to improve HSI classification accuracy. Haut *et al.* [26] incorporated visual attention-driven techniques into the ResNet to better represent the spectral–spatial information. Zheng *et al.* [27] used a full convolutional network (FCN)-based encoder–decoder structure and a fast patch-free global learning method to improve the convergence speed and accuracy of HSI classification. Although CNN-based backbone architectures can achieve state-of-the-art (SOTA) HSI classification performance, a number of critical issues still exist. For instance, the convolutional kernel has a single shape and a limited size. Since most of the land cover categories have irregular shapes, it is difficult for a single fixed square kernel to capture the complete feature information of objects. Meanwhile, the small kernel size also limits the CNN receptive field.

Recently, Transformer [28] architecture based on the self-attention mechanism has demonstrated strong potential to replace the standard CNN and has been regarded as a classical model in natural language processing (NLP) [29], [30]. The Transformer model has been also carried out in the computer vision community, and the representative work includes ViT [31], DeiT [32], and so on. The standard vision Transformer (ViT) treats an image as a sequence of nonoverlapping, fixed-size patches that are fed into the Transformer blocks after a linear embedding layer to model the long-range dependency. The Transformer has been attempted in a few studies on HSI classification. He *et al.* [33] proposed an HSI-BERT method for HSI classification using bidirectional encoder representation from Transformer and achieved better flexibility and generalization capability. Zhong *et al.* [34] integrated spectral attention and spatial attention modules, and proposed a novel spectral–spatial Transformer architecture. Hong *et al.* [35] added cross-layer skip connectivity to the Transformer for learning local spectral sequence information from adjacent bands. Yang *et al.* [36] proposed an HiT classification network by embedding the convolution-relevant modules into Transformer, allowing the extraction of slight spectral differences and the conveyance of information.

It should be noted that all of the above work is based on the direct translation of Transformer from NLP. However, in fact, there exist significant differences between NLP and RS. One of the differences is the size of basic elements. For instance, a word in NLP is a basic element with a fixed size, whereas, in RS, the basic element is a multiscale concept, which can be represented by pixels, objects, patches, or scenes. In the current Transformer-based models, processing units are all single size, which is certainly not conducive to many tasks in RS (e.g., object detection and semantic segmentation). The other difference is the number of basic elements. To be precise, an image contains many more pixels than words in a text paragraph. Therefore, it seems impossible to conduct the pixel-level dense prediction tasks for the RS images by directly borrowing the Transformer models from the NLP domain since the computational complexity of self-attention is quadratic to the image size. Given this, Liu *et al.* [37] proposed a generalized backbone network, i.e., the Swin Transformer (SwinT). Its computational complexity is linear to image size, and it also enables the construction of hierarchical feature maps so that more advanced techniques, such as feature pyramid network (FPN) [38], can be utilized. Considering the 3-D characteristics of HSI, this article aims to improve the original SwinT to the 3-D structure by proposing 3-D SwinT (3DSwinT), which can effectively reduce information loss and model the spatial–spectral dependencies.

On the other hand, the majority of current Transformer-based studies in RS are conducted in a supervised learning manner [33], [34], [35], which necessitates a large number of high-quality annotated samples and, therefore, is undoubtedly expensive and time-consuming [39]. Moreover, since RS images have very strong spatiotemporal heterogeneity and rich spectral information, it is difficult to annotate samples with wide coverage, multitemporal, multispectral, and multiresolution.

To address this issue, self-supervised learning (SSL) has been proposed and applied, which belongs to unsupervised learning and is intended to learn semantical representations from a large number of unlabeled images. In theory, images themselves should contain richer and more diverse information than the limited labels, which makes SSL easier to implement and more promising. Specifically, SSL methods first pretrain the feature extraction network to learn potential representations

from images and then fine-tune the pretrained network using a few labels in downstream tasks. SSL is an effective way to solve the "label starvation" problem for RS image deep learning. Since MoCo [40] achieved SOTA performance in the vision tasks, the contrastive learning [41] method has been gradually becoming the mainstream of SSL. Contrastive learning learns features by constructing positive and negative sample pairs, and its main idea is to minimize the distance between positive pairs and maximize the distance between negative ones. Contrastive learning has been successfully applied in the field of RS, and several notable examples include change detection [42], semantic segmentation [43], and scene classification [44]. It is also very promising to apply it to the HSI land cover classification. Xu *et al.* [45] proposed an end-to-end spectral–spatial unsupervised semantic feature extractor to learn the high-level semantic information from HSI and then adjusted the learned features with contrastive loss as the objective function.

In the field of CV, MoCo v3 [46] investigated several basic components for training ViT based on contrastive learning; Caron *et al.* [47] proposed a simple and efficient contrastive learning method called DINO and showed its synergy with ViT; and MoBY [48] combined MoCo v2 [49] and the BYOL [50] contrastive learning method, and made SwinT [37] as the backbone to evaluate its performance in downstream tasks. Comparatively speaking, in the field of RS, the research that involves or integrates Transformer and contrastive learning is scarce, and there are even fewer relevant studies for HSI classification.

In this study, we propose a hierarchical contrastive learning (HCL) framework to fully exploit the multiscale semantic information in the multiresolution feature maps. It should be noted that the multiscale features can also be effectively represented through the hierarchical feature construction ability of 3DSwinT. Typically, contrastive learning methods view a whole image as the learning target to extract image-level global representation, which, however, is ill-considered for downstream dense prediction tasks that necessitate pixel-level information. HSI classification is a pixel-by-pixel segmentation task, and therefore, extracting only global features will inevitably lead to the loss of many local details. To overcome this limitation, in our research, besides the global feature representation module, we also propose a multiscale local contrastive learning (MS-LCL) module to learn pixel-level representations by selecting geographically matched multiscale local regions from the multilevel feature maps output by 3DSwinT.

In summary, this article proposes a 3DSwinT-based HCL (3DSwinT-HCL) method for HSI land cover classification. To the best of our knowledge, this is the first time that contrastive learning and SwinT-based backbone have been combined for HSI classification. The main contributions of this article can be summarized as follows.

1) Proposed a novel 3-D architecture, called 3DSwinT for HSI classification.
2) Proposed a novel self-supervised contrastive learning method, namely, HCL. It consists of two components, i.e., the multiscale global contrastive learning (MS-GCL) module and an MS-LCL module.
3) The extensive experiments demonstrate the superiority of the proposed methods.

The remaining parts of this article are organized as follows. Section II presents the related work. The network architecture is described in detail in Section III. Data description, experimental setup, experimental results, and discussions are presented in Section IV. Section V concludes this article.

II. Related Work

A. Contrastive Representation Learning

SSL originated from NLP and is usually divided into two main categories: generative methods and contrastive methods [41], [51]. Generative methods are a pixel-level modeling approach, but they fail to establish spatial structure relationships since they focus on pixel details [52]. Contrastive methods utilize positive and negative samples to learn both the invariance of various augmented views of the same image and the ability to distinguish different images. Contrastive methods have now become the mainstream of SSL because of their superior performance and generalization ability. They usually employ the InfoNCE [51] loss function to learn representations, which requires a large number of negative examples, and the simplest and most straightforward way is to use large batches [53] or design memory banks to store all features [54]. The former is related to GPU capacity, and the latter demands a lot of memory.

To resolve the above problems, recent studies have attempted to improve the method while preserving the Siamese structure. In the BYOL model [50], negative samples were removed, and a momentum encoder, a prediction head, and gradient stopping strategies were adopted to avoid network collapse. SwAV [55] avoided collapse solutions by clustering, and SimSiam [56] verified that gradient stopping was the key to preventing network degradation. MoCo [40], [49] replaced the memory library with a queue dictionary, where features can be constantly updated to avoid memory consumption and the consistency of negative samples. In this study, we propose a novel contrastive learning method that incorporates queue design, momentum encoder, and prediction head, and more importantly, the proposed method has multiscale feature learning ability.

B. Self-Attention Mechanism and Transformer

The self-attention mechanism can model long-range dependency in sequence data and has been successfully applied in HSI classification. Fang *et al.* [57] introduced a spectral self-attention module into 3-D dilated convolution to enhance the distinguishability of spectral features. Sun *et al.* [58] proposed a spectral–spatial attention network to extract features from HSI cubes and mitigate the influence of irrelevant pixels. Zhu *et al.* [59] adopted spectral- and spatial-attention mechanisms on the basis of residual networks to adaptively select spectral bands and spatial information. However, although the above attention modules can achieve better performance, they are all constructed on CNN.

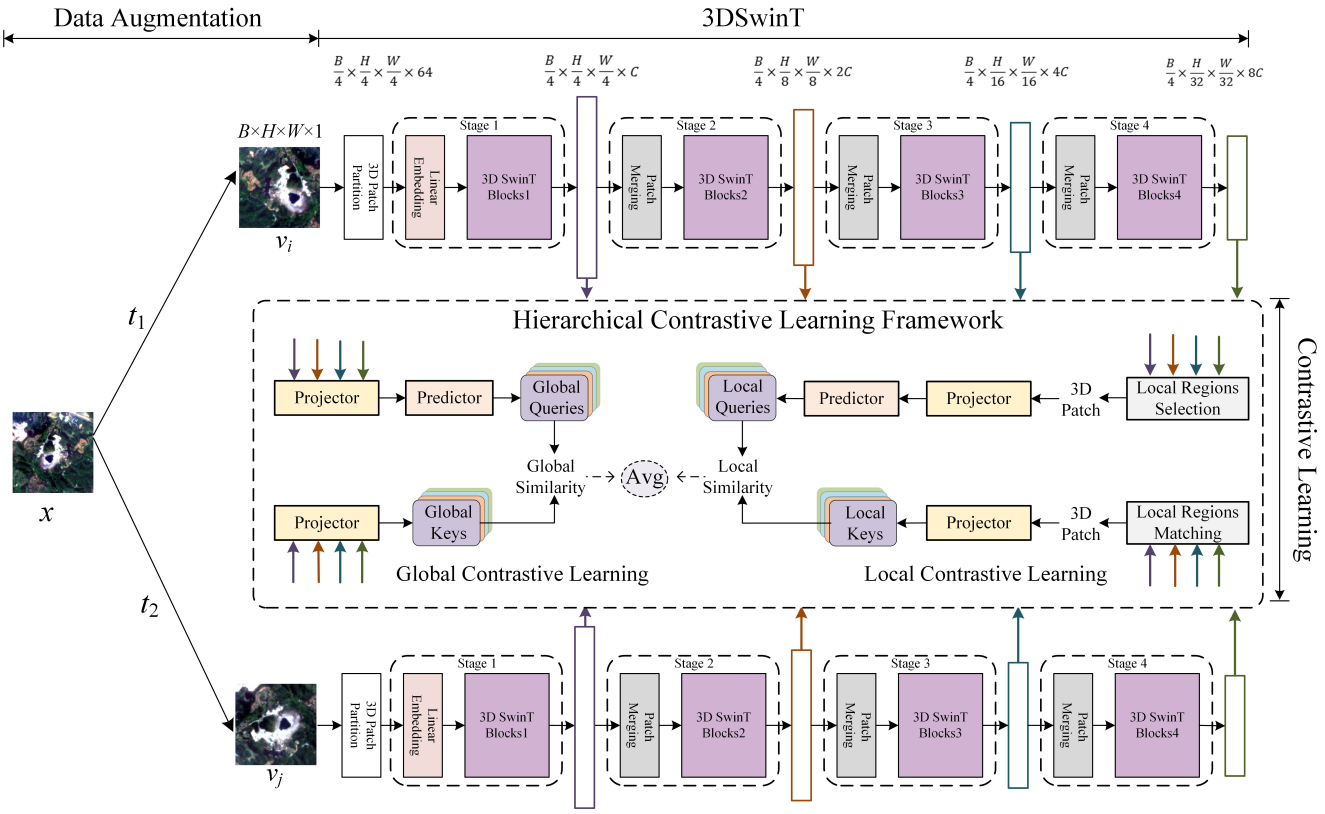

Fig. 1. Overall network architecture of the proposed 3DSwinT-HCL. 3DSwinT-HCL has a Siamese structure, consisting of data augmentation, 3DSwinT feature extraction, and HCL. Data augmentation includes geometric and color transformations with stochastic initialization parameters. 3DSwinT contains four stages, where the feature maps of different sizes are output. The hierarchical contrastive framework can learn multiscale representations and mine both image-level information and pixel-level information. (The figure does not show negative samples. For global contrastive learning, the negative samples are augmented views of other images from a batch. For local contrastive learning, the negative samples are local regions selected from other images.)

Transformer arose from machine translation and has become the dominant architecture in NLP. The Transformer is made up of an encoder and a decoder, both of which consist of multiple stacked self-attention blocks. ViT [31] is the pioneering work of Transformer in the field of CV, and its input is a series of nonoverlapping, medium-sized patches. Many related works followed ViT, such as DeiT [32], SwinT [37], PVT [60], and Twins [61]. These studies have improved ViT in terms of training strategies [32], hierarchical features [37], [60], computational complexity [37], and attention mechanisms [61]. In RS, ViT achieved a tradeoff between accuracy and efficiency in change detection, segmentation, and classification tasks [62], [63], [64]. However, ViT is still limited in dense prediction tasks or processing high-resolution images owing to its inherent structural nature.

C. Swin Transformer

Based on ViT [31], SwinT [37] introduced pyramid structure, locality, and translation invariance, and incorporated the sequence modeling capability and prior information of visual signals. Its computational complexity is linear to image size. SwinT v2 [65] further improved model capacity by proposing a postnormalization, log-spaced continuous position bias technique when training large models. PVT [60] had a pyramid structure similar to SwinTs, but its computational complexity remained quadratic to image size. PVT v2 [66] developed a linear spatial reduction attention mechanism to further reduce the complexity. Recently, some works have investigated the effect of fusing SwinT and convolutional networks [67]. It is worth noting that the SwinT-relevant studies are still scarce in RS tasks, especially the HSI classification. Gao *et al.* [68] combined the advantages of SwinT and CNN to construct a STransFuse model in order to extract coarse- and fine-grained features at various scales. Xu *et al.* [69] used SwinT as the backbone to model global relationships of images and accelerate network inference. This article extends SwinT to the 3-D structure in order to adapt to HSI classification.

III. Methods

A. Overall Network Architecture

SwinT considers hierarchy, locality, and translation invariance on the basis of ViT, on top of which our 3DSwinT further takes into account the 3-D characteristic of HSI. On the other hand, contrastive learning utilizes unlabeled samples with data augmentation strategies to pretrain the network and learn a large amount of potential semantic representations. Subsequently, the pretrained network can be fine-tuned with a few labels in downstream tasks. The downstream task of this article is pixel-level HSI semantic segmentation. Considering the 3-D nature of HSI and the multiple scales and sizes of ground

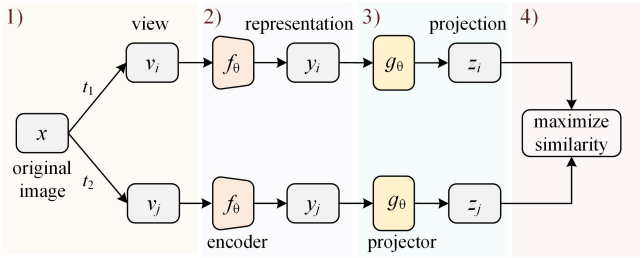

Fig. 2. Conventional contrastive learning framework. Referring to [53], the framework consists of data augmentation, representation extraction, nonlinear transformation, and contrastive loss.

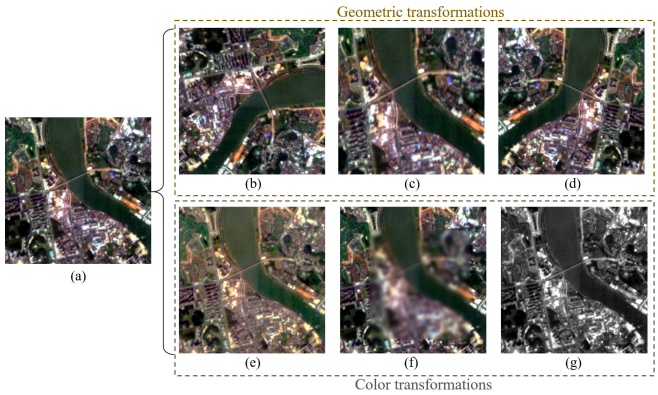

Fig. 3. Data augmentation. (a) Original image. (b) Rotation. (c) Crop and resize. (d) Flip. (e) Color distortion. (f) Random blur. (g) Gray. A stochastic initialization is adopted for these augmentation operations.

objects, we propose a novel HCL method based on 3DSwinT (3DSwinT-HCL). 3DSwinT-HCL is capable of simultaneously learning image- and pixel-level multiscale representations during the SSL pretraining phase, which is beneficial for better transfer to the downstream dense prediction task.

The overall network architecture of 3DSwinT-HCL is shown in Fig. 1, which includes the following steps.

1) For an input x, two augmentation views (v_i and v_j) are first generated by stochastic data augmentation strategies.
2) The generated views are then imported to the Siamese 3DSwinT networks for feature extraction. The two 3DSwinT branches are called online and target encoders, respectively. Each 3DSwinT consists of four stages, yielding feature maps of various sizes.
3) Finally, multiscale feature maps of the two branches are fed to the proposed HCL framework for multiscale global and local contrastive learning. 3DSwinT is capable of extracting the rich spatial and spectral information of HSI, and furthermore, the multiscale and global–local learning of HCL can adequately take into account the characteristics of the ground objects for the dense prediction tasks.

B. Contrastive Learning

The idea of contrastive learning is to maximize the similarity between positive pairs and minimize the similarity between negative ones. It is generally designed in the form of a Siamese network [70], whose general framework is presented in Fig. 2, containing four main components.

1) Data Augmentation: This module plays an important role in self-supervised contrastive learning and aims to apply a series of random transformations to a batch of input data in order to construct labels for contrast. Labels can be divided into anchors, positive samples, and negative samples. For an image, data augmentation can build noise-free representations and generate positive samples with similar features, whereas, in contrast, the augmented views of different images are treated as negative samples. In this study, we adopt geometric and color transformations to produce diverse samples. As shown in Fig. 3, specifically, the geometric transformations include random cropping, scaling, rotation, and flipping, and the color transformations consist of color distortions (brightness, contrast, saturation, and hue), blurring, and graying. Each input x results in two related views by the stochastic data transformations, t_1 and t_2, i.e., $v_i = t_1(x)$ and $v_j = t_2(x)$, where v_i can be referred to as the anchor, v_j is its matching positive sample, and augmented views of different images are the negative ones.

Negative samples are indispensable for contrastive learning, without which the network may collapse. There are two methods for constructing and updating negative samples.

Method 1: Negative samples are the augmented views of other images apart from the anchor from the same batch and are updated end-to-end by back propagation [53]. In this way, for a batch containing N images, there are $2N$ samples after data augmentation. Given an anchor, there is only one matching positive sample, and the remaining $2(N-1)$ samples are negative ones.

Method 2: The negative samples of each batch are stored in a large dictionary, which is maintained as a queue and updated by the momentum encoder [40]. Continuous replacement between new and old samples in the queue ensures the consistency of negative samples. The queue size can be viewed as a hyperparameter since it is decoupled from the batch size. Therefore, this method can produce more negative samples. In this article, we utilize Method 2 to generate negative samples.

2) Representation Extraction: The neural network encoder f_θ can be used for information extraction and feature transformation in order to build representations for downstream tasks. f_θ allows multiple options, such as convolutional neural or Transformer-based networks. This article proposed a 3DSwinT backbone to generate representations $y_i = \text{3DSwinT}(v_i)$ and $y_j = \text{3DSwinT}(v_j)$, where an adaptive average pooling operation is required for y_i and y_j.

3) Nonlinear Transformation: A nonlinear projection head g_θ further transforms the extracted representation to the projection layer z, where the loss value is calculated. g_θ consists of a multilayer perceptron (MLP) with one hidden layer, yielding $z_i = W^{(2)}\sigma(W^{(1)}y_i)$ and $z_j = W^{(2)}\sigma(W^{(1)}y_j)$, where σ is the rectified linear unit (ReLU). The module was first introduced by SimCLR [53], can effectively avoid information loss, and, hence, improve the effectiveness of representations.

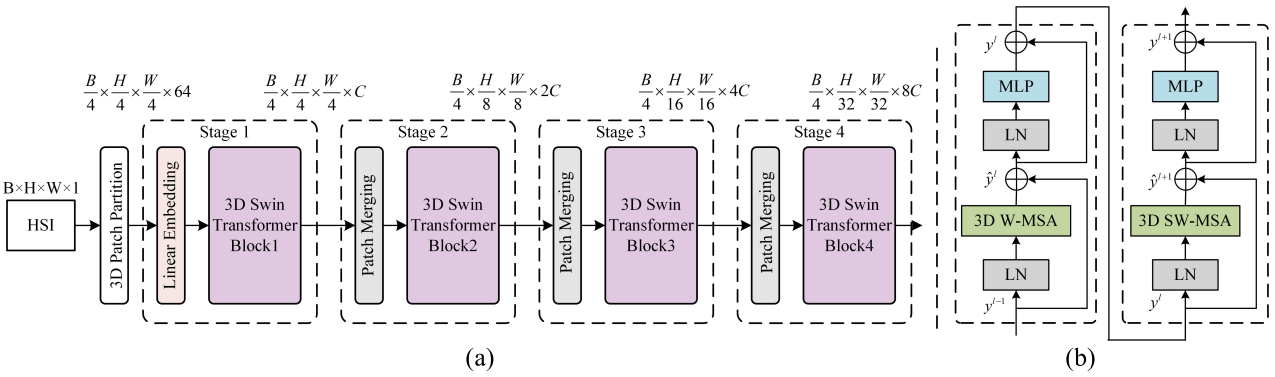

Fig. 4. Overall architecture of 3DSwinT. (a) Network architecture. (b) Two consecutive 3DSwinT blocks. 3DSwinT is made up of four stages, with each outputting the feature maps of different scales. 3DSwinT blocks contain LN, 3DW-MSA, MLP, and residual connection.

4) Contrastive Loss: As the objective function of self-supervised contrastive learning, the contrastive loss primarily aims to train the encoder network. For a dataset of $\{x_k\}$, and a given anchor x_i, the contrastive loss requires minimizing the distance between x_i and its positive sample x_j, and maximizing the distance between x_i and the negative samples $\{x_k\}_{k\neq j}$. Based on the commonly used InfoNCE [51], we define the loss function for a positive pair of $(i,\ j)$ as

$$\ell_{(v_i,v_j)} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\exp(\text{sim}(z_i, z_j)/\tau) + \sum_{z\in\Lambda^-} \exp(\text{sim}(z_i, z)/\tau)} \tag{1}$$

where τ is the temperature parameter, sim denotes the similarity between samples, which is often measured by cosine similarity, and Λ^- represents all the negative samples. For a batch containing N samples, we can obtain the final loss value across all positive pairs, i.e., $(i,\ j)$ and $(j,\ i)$

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}(\ell(v_i, v_j) + \ell(v_j, v_i)). \tag{2}$$

From (1), contrastive learning needs to simultaneously construct both positive and negative pairs. If there are only positive pairs (without negative pairs), the model tends to encode all samples into the same feature, leading to degenerate solutions. Conversely, if there are no positive samples, the model lacks clustering ability. Existing studies [40], [53] show that more negative pairs can lead to stronger learning capabilities. This is because more negative pairs can describe the underlying distribution more effectively, thus optimizing the training direction and accelerating convergence. In addition, negative samples are preferably close to the positive ones but with different labels. Such samples are called hard negatives.

C. 3DSwinT

SwinT can construct multiscale feature maps by continuously fusing neighboring patches and the window partition mechanism, and its computational complexity is linear to image size, which is beneficial for dense prediction tasks and high-resolution images. In this study, we extend SwinT to a 3-D structure, i.e., 3DSwinT, to accommodate the 3-D properties of HSI and capture its rich spatial and spectral information. Fig. 4 depicts the architecture of 3DSwinT. Compared to SwinT, the improvements made are summarized in the following aspects.

1) We define each HSI as $B \times H \times W \times 1$, where B is the number of HSI bands, and H and W denote the height and the width of the image, respectively.
2) In the patch partition module, SwinT splits the input into $(H/4) \times (W/4)$ patches with a size of 4×4. In contrast, our proposed 3DSwinT takes a 3-D cube $(4 \times 4 \times 4)$ as the basic processing unit, leading to a total of $(B/4) \times (H/4) \times (W/4)$ patches, with a feature dimension of 64, and then, a linear embedding layer projects these patches to an arbitrary dimension of C. The neighboring patches are merged in the subsequent patch merging phase, and the spatial size of the patches becomes 4, 8, 16, ... in sequence while keeping the spectral domain constant.
3) The difference between 3DSwinT blocks and SwinT blocks lies in the window-based multihead self-attention (W-MSA) mechanism. We add the spectral domain to W-MSA, yielding 3-D W-MSA (3DW-MSA), by considering the window partitioning and shifting mechanism (as shown in Fig. 5). SwinT adopts 2-D windows of size $M \times M$ to divide input patches evenly, while 3DSwinT utilizes 3-D windows of size $P \times M \times M$. In addition, we refine the original window shifting mechanism by moving $(P/2,\ M/2,\ M/2)$ patches along the spectral, height, and width dimensions in the next block [see Fig. 5(b)] in order to strengthen information interaction between windows.

3DSwinT consists of four stages. Each stage includes a patch merging module and a series of 3DSwinT blocks (except for Stage 1). As mentioned above, the patch merging module only downsamples the spatial dimension (not the spectral dimension) to concatenate the neighboring 2×2 patches into a large patch. This means that the size of patches becomes four times that of the original, and the number becomes one-quarter of the original. Meanwhile, a linear layer is used to project the concatenated dimension to half of its original size. Finally, the 3DSwinT blocks are utilized to extract the self-attention information. This process does not change the

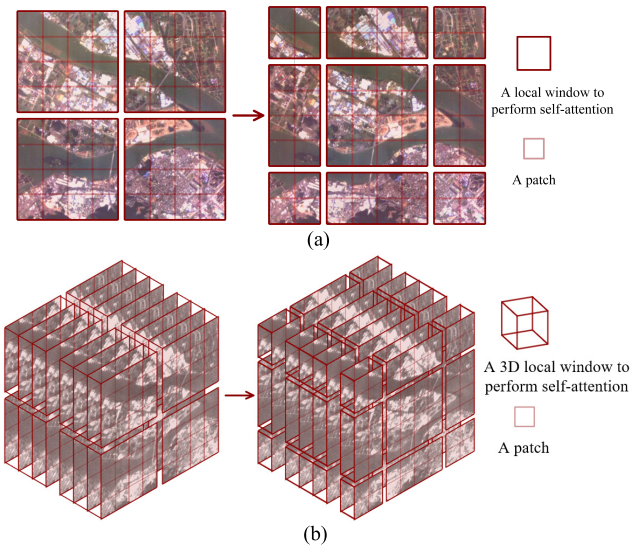

Fig. 5. Window partitioning and shifting of (a) SwinT and (b) 3DSwinT. Compared to SwinT, 3DSwinT uses the 3-D windows to divide input patches evenly and moves patches along the spectral, height, and width axes in the next block to strengthen information interaction between windows. (a) Window partitioning and shifting mechanism of SwinT. (Left) Regular window partition. (Right) Window shifting mechanism. (b) Window partitioning and shifting mechanism of 3DSwinT. (Left) Regular window partition. (Right) Window shifting mechanism.

input resolution. In this way, for Stage 1, the size of the feature map is $(B/4) \times (H/4) \times (W/4) \times C$, Stage 2 is $(B/4) \times (H/8) \times (W/8) \times 2C$, and so on for other stages.

In comparison to SwinT blocks, we employ 3DW-MSA to extract both spectral and spatial sequence information. All other components of 3DSwinT blocks are kept the same as SwinT, such as MLP, layer normalization (LN), and residual connection. Fig. 4(b) depicts two adjacent 3DSwinT blocks within each stage, which can be represented by following the equation:

$$\begin{aligned} \hat{y}^{l} &= \text{3D W-MSA}\left(\text{LN}\left(y^{l-1}\right)\right) + y^{l-1} \\ y^{l} &= \text{MLP}\left(\text{LN}\left(\hat{y}^{l}\right)\right) + \hat{y}^{l} \\ \hat{y}^{l+1} &= \text{3D SW-MSA}\left(\text{LN}\left(y^{l}\right)\right) + y^{l} \\ y^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{y}^{l+1}\right)\right) + \hat{y}^{l+1} \end{aligned} \tag{3}$$

where 3DW-MSA and 3-D SW-MSA represent the 3-D window-based and shifted W-MSA mechanisms, respectively, and $\hat{y}^{l}$ and y^{l} are the outputs of 3-D (S)W-MSA and MLP in block l, respectively.

D. Hierarchical Contrastive Learning

Conventional contrastive learning methods usually feed data-augmented images into the Siamese network to construct representations and then perform contrastive learning in the representation space. Notice that current contrastive learning studies can only represent single-scale content but also, in fact, objects have complex and variable scales and sizes. Consequently, single-scale contrastive learning methods are not sufficient for the semantic segmentation task. Given this, we propose an HCL method that utilizes the multiresolution feature maps output by 3DSwinT to mine multiscale semantic

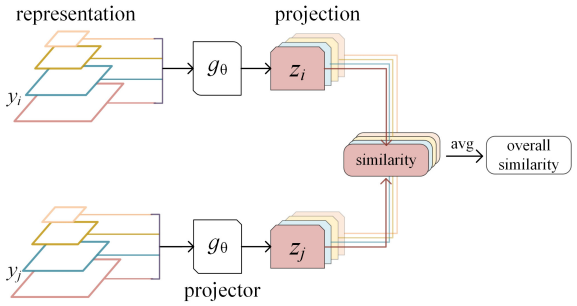

Fig. 6. Our HCL framework. The multiscale feature maps output by 3DSwinT are fed into the multiscale contrastive framework. Each level carries out contrastive representation learning at different scales in parallel, and the learning results of all levels are then fused to obtain the overall similarity of the network.

information. As shown in Fig. 1, the multilevel feature maps derived from both 3DSwinT branches are fed into the HCL framework. Specifically, the proposed method takes multiscale information extraction as the overall framework and includes both global and local contrastive learning modules to learn image- and pixel-level hierarchical representations at the same time. The following are the details.

Compared

1) Hierarchical Contrastive Learning Framework: with existing contrastive learning frameworks, our proposed HCL can simultaneously learn multiscale representations during self-supervised pretraining. Specifically, each level carries out contrastive representation learning at different scales in parallel, and the learning results of all levels are fused to obtain the overall similarity of the network, as shown in Fig. 6. With the hierarchical learning framework, both global information and local information are extracted to achieve image- and pixel-level multiscale representations.

2) Multiscale Global Contrastive Learning Module: This module conducts image-level hierarchical representation learning. Specifically, the output of 3DSwinT [denoted as $e(\cdot)$] at Stage s ($s = 1, 2, \ldots$ denotes different scales) is entirely fed into the module. Besides the projection head $g(\cdot)$, the module additionally introduces a prediction head $h(\cdot)$, mainly to prevent degenerate solutions, and it also consists of an MLP and a single hidden layer (ReLU). The module employs an asymmetric structure, and only the online encoder contains $h(\cdot)$. Therefore, for any input image x, we can obtain

$$\begin{aligned} z_i^s &= h\left(g\left(\text{Avg}\left(e^s(t_1(x))\right)\right)\right) \\ z_j^s &= g\left(\text{Avg}\left(e^s(t_2(x))\right)\right) \end{aligned} \tag{4}$$

where $e^s(\cdot)$ stands for Stage s of the encoder $e(\cdot)$. Avg is the adaptive average pooling. z_i^s and $z_j^s \in R^D$, for the calculation of global similarity at different scales.

The idea of contrastive learning is to learn similar/dissimilar representations from positive and negative pairs and, thus, can be characterized as a dictionary lookup task [40]. Following [40], we store negative features of each batch during the training in a dictionary, which is maintained as a queue and updated by the momentum encoder in order to ensure the consistency of features. The queue size can be much larger

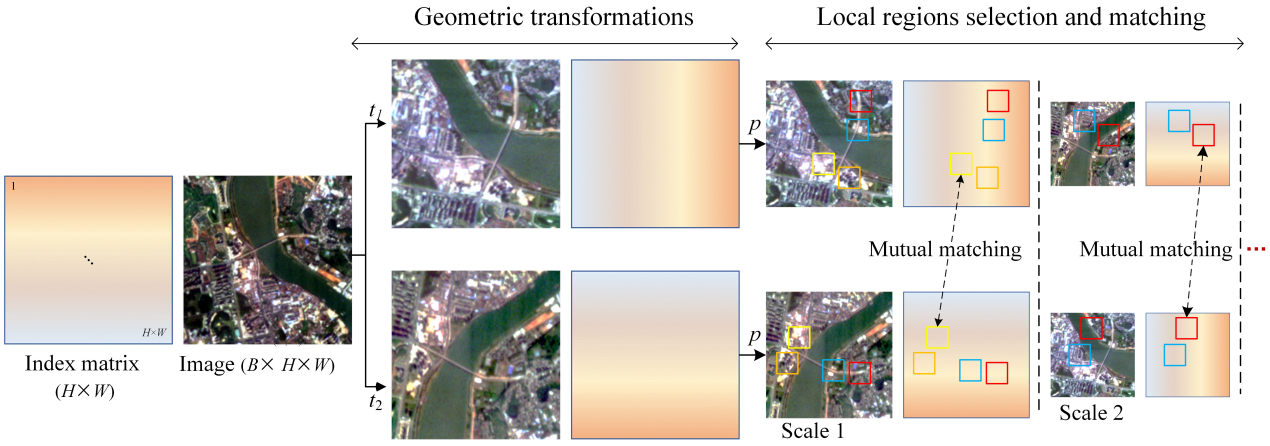

Fig. 7. Selection and matching of multiscale local regions. The index matrix is used to select the matching positive sample pairs, which has the same size as the spatial size of the image and is spatially transformed along with the images. We use maximum pooling to make the matrix the same size as the feature map at each stage. Finally, we select a number of local regions in terms of the same pixel value of the two index matrices for each scale.

than the batch size since it is independent of the batch, so as to generate sufficient and diverse negative samples.

Query and key are two key elements in the dictionary lookup task. Each query q matches a positive key k_+ in the dictionary to constitute a positive pair, i.e., two augmented versions of the same image. On the other hand, the remaining features form the negative keys $\{k_-\}$. The contrastive function is dedicated to maximizing the similarity between q and k_+ while distinguishing q and $\{k_-\}$. According to InfoNCE, we define the hierarchical global loss for an input batch containing N samples as

$$\mathcal{L}_G^s = \frac{1}{2N}\sum_{1}^{N}\left(\ell_g^s(v_i, v_j) + \ell_g^s(v_j, v_i)\right) \tag{5}$$

with

$$\ell_g^s(v_i, v_j) = -\log\frac{\exp\left(z_i^s \cdot z_j^s/\tau\right)}{\exp\left(z_i^s \cdot z_j^s/\tau\right) + \sum_{k=1}^{K}\exp\left(z_i^s \cdot z_k^s/\tau\right)} \tag{6}$$

where we employ dot products to measure the similarity between samples. $\sum_{k=1}^{K} z_k^s$ denotes all global negative samples for each scale in the dictionary, and we set the negative sample size to K since it is decoupled from the batch size. In this study, s is set to 1, 2, 3, 4, and hence, we can obtain four global losses representing different scales: $\mathcal{L}_G^1$, $\mathcal{L}_G^2$, $\mathcal{L}_G^3$, and $\mathcal{L}_G^4$.

3) Multiscale Local Contrastive Learning Module: As a pixel-level segmentation task, HSI classification demands local detail information, which cannot be represented effectively in the global representation module. Therefore, we propose an MS-LCL module to focus on local regions in order to learn the pixel-level representations with multiscale properties. Unlike global contrastive learning, local contrastive learning begins with the selection and matching of local regions [denoted as $m(\cdot)$], which is dealt with by choosing geographically matched local regions based on the multilevel feature maps output by the Siamese 3DSwinT. The specific steps are introduced as follows.

Step 1: To ensure that the selected local regions are geographically aligned, we employ a 2-D index matrix with the same spatial size as the image to record the pixel positions of the image. Since pixel positions are disrupted and cannot be aligned after random data augmentation, index matrices need to be transformed along with images, specifically geometric transformation, i.e., random cropping, scaling, rotation, and flipping.

Step 2: As previously stated, 3DSwinT outputs feature maps of different sizes at each stage. Given an image with a size $B \times H \times W$, the feature map constructed by Stage s of 3DSwinT can be expressed as $(B/4) \times (H/(4 \times 2^{s-1})) \times (W/(4 \times 2^{s-1})) \times (2^{s-1} \times C)$. Therefore, we reshape the index matrix to the same size as the feature map at each scale in order to guarantee their spatial matching. We achieve this requirement through the pooling operator $[p(\cdot)]$.

Step 3: In this way, multilevel index matrices that record precise pixel locations can be obtained. Subsequently, we can select a number of matching pixels in terms of the same pixel value of the two index matrices for each scale, where the pixels correspond to local regions of the original images.

The implementation of the above steps is demonstrated in Fig. 7. Therefore, for the module, we can obtain

$$\begin{aligned} u_i^s &= h\left(g\left(m\left(\text{Avg}\left(e^s(x)\right)\right)\right)\right) \\ u_j^s &= g\left(m\left(\text{Avg}\left(e^s(x)\right)\right)\right) \end{aligned} \tag{7}$$

where u_i^s and $u_j^s \in R^D$ for the calculation of local similarity at different scales.

Analogously to the global contrastive loss, the hierarchical local loss can be written as

$$\mathcal{L}_L^s = \frac{1}{2N_L^s}\sum_{1}^{N_L^s}\left(\ell_l^s(v_i, v_j) + \ell_l^s(v_j, v_i)\right) \tag{8}$$

with

$$\ell_l^s(v_i, v_j) = -\log \frac{\exp\left(u_i^s \cdot u_j^s/\tau\right)}{\exp\left(u_i^s \cdot u_j^s/\tau\right) + \sum_{k=1}^{K_L^s} \exp\left(u_i^s \cdot u_k^s/\tau\right)} \tag{9}$$

where N_L^s indicates the total number of matching local regions at each scale chosen from a batch containing N samples, i.e., $N_L^s = N \times n_m^s$, with n_m^s being how many local regions are selected from a sample. $\sum_{k=1}^{K_L^s} u_k^s$ are all local negative samples for each scale in the dictionary, i.e., $K_L^s = K \times n_m^s$. In this way, (8) yields four local losses representing different scales, namely, $\mathcal{L}_L^1$, $\mathcal{L}_L^2$, $\mathcal{L}_L^3$, and $\mathcal{L}_L^4$.

4) Total Network Loss: The total network loss can be obtained based on the above four-level global and local losses

$$\mathcal{L} = \frac{1}{4} \sum_{s=1}^{4} \left(\lambda \times \mathcal{L}_G^s + (1-\lambda) \times \mathcal{L}_L^s\right). \tag{10}$$

IV. Results and Discussion

A. Data

1) Zhuhai-1 Hyperspectral Data (OHS): The Zhuhai-1 hyperspectral constellation consists of ten Orbita hyperspectral micro–nano satellites, which are operated and managed by Orbita Corporation, China. For the first time, OHS achieves a hyperspectral satellite network for a rapid response to earth observation. The imaging resolution of each satellite is 10 m, and there are 256 spectral bands ranging from 400 to 1000 nm. Orbita Aerospace provides 32 spectral bands that are selected from the 256 channels according to the users' needs. In this research, we chose the default 32 bands, whose central wavelengths range from 466 to 940 nm. Specifically, in the experiments, 400 images with a size of 224 × 224 are used for algorithm testing, and four land cover classes, vegetation, building, bare land, and water are included by considering the land cover characteristics of the study area and the spatial resolution of OHS. The images are divided into self-supervised pretraining and test sets with a ratio of 9:1, and 10% of pretraining tests are for fine-tuning the downstream task.

2) Six Widely Used Hyperspectral Datasets: We choose six existing hyperspectral datasets to evaluate the transfer capability of the 3DSwinT-HCL method, including Indian P,[1] Salinas,[1] Botswana,[1] Pavia U,[1] Pavia C,[1] and DFC2018[2] (see Table I). The partitioning of the training and testing sets follows the providers' recommendations. The number of training and testing samples for the six hyperspectral datasets is shown in Table II.

First, we use OHS data for self-supervised pretraining, and then, the pretrained model is fine-tuned with a few labels. Finally, we transfer the pretrained model to the six HSIs for classification.

B. Experimental Setup

1) Evaluation Metrics: OA and Kappa are used to quantitatively evaluate the performance of different algorithms.

[1]http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[2]http:i/hyperspectral.ee.uh.edu/?page_id=1075

TABLE I
HYPERSPECTRAL DATASETS USED IN THIS ARTICLE

Dataset	Spatial Resolution (m)	Size (pixels)	Spectral Bands	Spectral Range (um)	Classes
OHS	10	224×224	32	0.4-1	4
Indian P	20	145×145	200	0.4-2.5	16
Salinas	3.7	512×217	204	0.4-2.5	16
Botswana	30	256×1476	145	0.4-2.5	14
Pavia U	1.3	610×340	103	0.43-0.85	9
Pavia C	1.3	1096×715	102	0.43-0.85	9
DFC2018	1.0	4786×1202	48	0.38-1.05	20

TABLE II
SAMPLE SIZE OF SIX COMMONLY USED HYPERSPECTRAL DATASETS

Dataset	Training	Testing
Indian P	5538	4711
Salinas	17511	36618
Botswana	1394	1854
Pavia U	2774	40002
Pavia C	53933	94219
DFC2018	248338	256374

2) Comparison With SOTA Methods: The proposed 3DSwinT-HCL is compared with other contrastive learning methods, including MoCov3 [46], DINO [47], MOBY [48], and SiT [71]. Furthermore, the self-supervised pretrained 3DSwinT-HCL is transferred to the six commonly used hyperspectral datasets to assess its transferability.

3) Implementation Details: In the SSL phase, we train all the self-supervised models for 500 epochs using the AdamW optimizer, with a batch size of 16. The initial learning rate is 0.001, with a cosine decay schedule. The momentum and weight decays are 0.9 and 0.05, respectively. For the MS-LCL module of our 3DSwinT-HCL, the number and size of local areas are set to $8 \times 4 \times 4$, $4 \times 8 \times 8$, $2 \times 16 \times 16$, and $1 \times 32 \times 32$ at the four stages, respectively. In the fine-tuning phase, we only use 10% labels of self-supervised samples to fine-tune the network, with a learning rate of 0.0005, a batch size of 26, and CrossEntropy as the loss function.

C. Experimental Results

In this section, we first compare the 3DSwinT-HCL method with other methods combining contrastive learning and Transformer based on the OHS dataset, and then, we analyze the difference between self-supervised and supervised learning. Finally, considering the large differences in spectrum and resolution for different hyperspectral sensors, it is interesting to test whether 3DSwinT-HCL can be effectively transferred to other hyperspectral datasets.

1) Comparison With Other Contrastive Learning Methods: Results are shown in Table III.

From Table III, we can see that our method obtains the highest accuracy, with an OA of 80.15% and a Kappa of 0.70, which is the only model with an OA greater than 80%. In contrast to MOBY, which also adopts SwinT as the

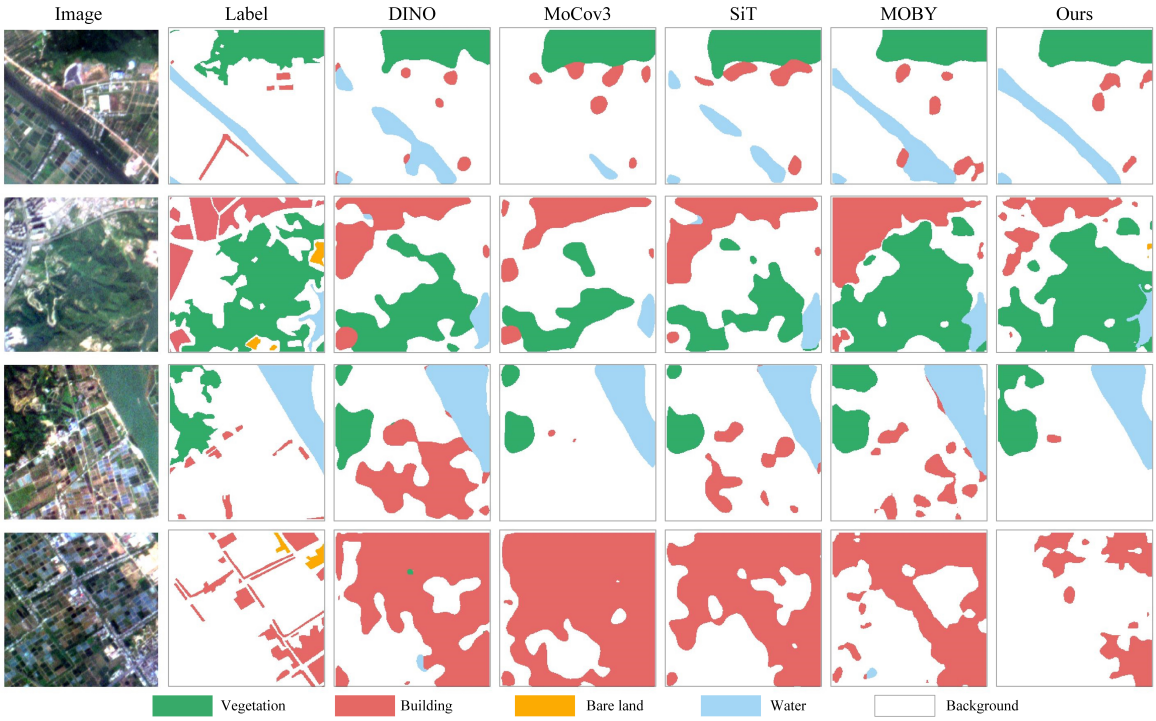

Fig. 8. Visualization of classification results on the OHS dataset. Our method achieves the highest classification accuracy, presents fewer false alarms, and is better at describing the completeness and details of objects.

TABLE III
COMPARISON BETWEEN SELF-SUPERVISED METHODS (WITH 10% LABELS FOR FINE-TUNING)

Method	Arch.	OA(%)	Kappa	params(M)
MoCo v3	ViT	75.20	0.62	42.82
SiT	ViT	75.68	0.64	27.86
DINO	ViT	74.60	0.63	30.16
MOBY	SwinT	78.83	0.69	17.55
Ours	3DSwinT	80.15	0.70	28.90

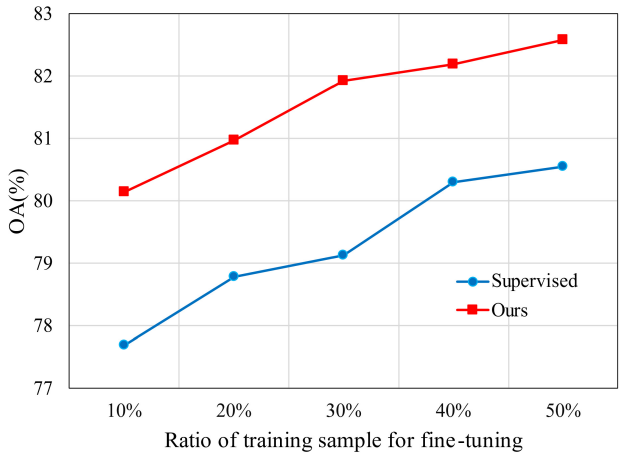

Fig. 9. Comparison of classification accuracy with different sample ratios. The accuracy of both self-supervised and supervised learning gradually improves as the ratio of labels increases. It can be seen that our method outperforms the supervised learning by an average of 2.3%.

backbone, 3DSwinT-HCL obtains better accuracy by 1.3%. On the one hand, the existing contrastive learning approaches only focus on the single-scale feature representation but do not take into account the multiscale characteristics of the land cover classes. On the other hand, their backbones seem inadequate to extract rich spectral and spatial information from HSI. Therefore, they fail to achieve optimal results in the semantic segmentation task. In contrast, our proposed 3DSwinT-HCL can adaptively learn the multiscale semantic information and fully model the spatial–spectral dependencies during self-supervised pretraining. In addition, 3DSwinT and HCL demonstrate good complementarity. In this way, satisfactory performance can be obtained with only a few labels in the downstream task. In addition, the MS-LCL module embedded in the hierarchical framework can additionally focus on the local details of objects and learn pixel-level representations, which are more beneficial for dense prediction tasks.

Moreover, we visualize a portion of the classification results in Fig. 8. It can be seen that the other methods produce a large number of false alarms, but our method can reduce this kind of error and has fewer misclassifications. Comparatively speaking, ours is better at depicting the completeness and details of objects. Inevitably, however, it can be observed that our method has drawbacks, especially in the boundary areas. However, this can be acceptable given the very limited amount of samples used in the SSL. In summary, in the above experiment, our proposed 3DSwinT-HCL method shows great potential for HSI classification.

2) Self-Supervised Fine-Tuning and Supervised Learning: The idea of SSL is to pretrain the network on a large amount

of unlabeled data and then fine-tune it using a few labels in the downstream target, whereas supervised learning directly utilizes a large number of labels for feature learning. In this subsection, using 3DSwinT as the backbone, different ratios of labels are utilized to fine-tune the self-supervised model, and meanwhile, pure supervised learning is also performed in order to verify the effectiveness of our method. From Fig. 9, it is shown that, as the ratio of labels increases, the accuracy of both self-supervised and supervised learning gradually improves, and the 3DSwinT-HCL method consistently outperforms supervised learning. In terms of OA, our method surpasses supervised learning by an average of 2.3% with the same sample ratio and 2.8% with a 30% ratio. These results demonstrate that 3DSwinT-HCL pretraining is able to adaptively explore more latent features and is superior for HSI classification compared to supervised learning. To summarize, our results demonstrate the great potential and efficiency of SSL for HSI semantic segmentation.

D. Transfer to Other HSI Datasets

Transferability is tested in this section to evaluate whether the proposed contrastive learning method can perform well on other HSI datasets. The 3DSwinT model is pretrained by the HCL self-supervised method based on the OHS dataset and then transferred to the six commonly used hyperspectral datasets. It is a great challenge, as the images from different sensors have large differences in landscape characteristics, spatial resolution, and spectral channels. Specifically, OHS data has 32 spectral bands ranging from 0.46 to 0.94 μm, with an average spectral resolution of 2.5 nm and a spatial resolution of 10 m. In contrast, the minimum and maximum numbers of spectral channels for the six HSI datasets are 48 and 204, respectively, and the wavelength range is from 0.38 to 2.5 μm with a spatial resolution of 1–30 m. Considering the great differences between OHS and these datasets, the transfer capability of 3DSwinT-HCL is tested in the following three scenarios.

1) *All_Random:* We train the network from scratch using all the spectral bands for each HSI dataset. The network is randomly initialized without self-supervised information from 3DSwinT-HCL.
2) *Sub_Random:* For each HSI dataset, 32 bands are chosen according to the central wavelength of the OHS images, and the network is then also trained from scratch.
3) *Sub_HCL:* Unlike 2), the network is not randomly initialized but rather fine-tuned by the 3DSwinT-HCL's pretraining model.

For DFC2018 and Pavia C, 50 epochs are trained due to their relatively large training sample sizes, and the epoch for other datasets is set to 100. For all datasets, we set the batch size to 100, with a learning rate of 0.0005, and AdamW is used as the optimizer.

According to the results (see Fig. 10), Sub_HCL is always superior to Sub_Random in all the datasets and outperforms All_Random in the majority of them (except for Pavia C). It can be seen that, despite the significant spatial and spectral differences between OHS and other HSI datasets, the model

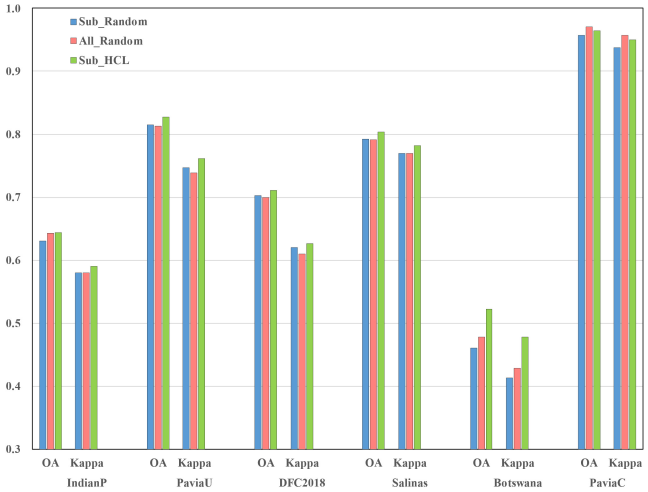

Fig. 10. Classification results of the SSL transferred to other commonly used hyperspectral datasets. Sub_HCL is always superior to Sub_Random for all the datasets and outperforms All_Random in most of the cases (except for Pavia C).

TABLE IV
ABLATION EXPERIMENTS OF 3DSWINT-HCL

3DSwinT	MS-GCL	MS-LCL	OA(%)	Kappa
×`	×	×	78.83	0.69
√	×	×	79.27	0.68
√	√	×	79.93	0.70
√	×	√	78.65	0.67
√	√	√	80.15	0.70

with 3DSwinT-HCL pretraining is always capable of achieving satisfactory results when transferring to other datasets. The classification accuracy of all datasets can be improved to some extent, regardless of whether the resolution of the datasets is higher or lower than 10 m of OHS. Therefore, it can be stated that our proposed 3DSwinT-HCL method has promising transfer ability, and the information derived from the SSL is beneficial for other datasets with different features.

E. Ablation Experiments

This article proposes the 3DSwinT-HCL method, and the technical contributions include the 3DSwinT backbone to extract both spatial and spectral features of HSI; the HCL framework for multiscale representation learning, including the MS-GCL module to exploit image-level information; and the MS-LCL module to capture local details. In this section, we conduct ablation experiments to investigate the importance of these components. Experiments (see Table IV) show that, when not using our proposed components (including 3DSwinT, MS-GCL, and MS-LCL), i.e., only the single-scale global contrastive learning is carried out with 2DSwinT as the backbone, the OA is 78.83%. When the backbone (2DSwinT) is replaced by 3DSwinT, the OA increases to 79.27%, with an improvement of 0.44%. Furthermore, when the global contrastive learning is embedded into the proposed HCL framework, i.e., MS-GCL, OA is raised to 79.93%, with an

TABLE V
COMPARISON WITH SOTA CNN NETWORKS

Backbone	Pretrain	OA	Kappa
FPGA	-	0.80	0.70
SSDGL	-	0.78	0.67
3DSwinT	-	0.80	0.70
3DSwinT	ImageNet-1k	0.83	0.75

TABLE VI
COMPARISON WITH SOTA SUPERVISION METHODS

Method	OA	Kappa
3D-FCN	0.77	0.66
SSRN	0.79	0.69
FPGA	0.79	0.69
SSDGL	0.78	0.67
Ours	0.80	0.70

TABLE VII
IMPACT OF DATA AUGMENTATIONS ON CONTRASTIVE LEARNING

Geometric Trans.	Color Trans.	OA	Kappa
√	×	77.94	0.66
×	√	77.60	0.65
√	√	80.15	0.70

TABLE VIII
EFFECTS OF GEOMETRIC TRANSFORMATIONS ON CONTRASTIVE LEARNING

W/o	OA(%)	Kappa
Rotation	79.73	0.69
Crop	79.06	0.68
Flip	79.53	0.69
--	80.15	0.70

increment of 0.66%. On the other hand, when the global learning module is removed and only MS-LCL is considered, the OA is 78.65%, indicating that global contrastive learning is indispensable in the proposed framework. In detail, the results show that global contrastive learning is the key module of the contrastive prediction task, while local contrastive learning can be viewed as its complementary module. Finally, OA can be increased to 80.15% while using all the proposed components.

On the one hand, the 2-D network may cause spectral information loss, while 3DSwinT mitigates this issue to some extent. On the other hand, compared to the single-scale contrastive structure of conventional methods, our HCL takes into account the fact that ground targets vary in scale and size, and hence, allows for multiscale representation learning. This consideration enables the network parameters to be updated in a more rational direction. Moreover, from another perspective, the pixel-level information learning ability of MS-LCL is beneficial for the dense prediction task.

F. Discussions

The above experimental results demonstrate the effectiveness of the proposed HCL self-supervised method and the complementarity with the 3DSwinT backbone. In this section, we conduct additional experiments and comparisons to further validate the methods presented in this article, including: 1) comparison with SOTA CNNs to verify the benefits of the proposed 3DSwinT for HSI land cover classification; 2) comparison with SOTA supervised methods to show the efficiency and potential of the HCL self-supervised method; and 3) the impact of data augmentations on contrastive learning performance.

1) Comparison With SOTA CNNs: We compare two SOTA CNNs that are specifically designed for HSI classification, i.e., FPFA [27] and SSDGL [72]. In order to conduct a fair comparison, these two SOTA CNN models are also implemented under the proposed HCL framework. Experiments demonstrate (see Table V) that our proposed 3DSwinT outperforms SSDGL and is comparable to FPGA in HSI classification under the same training setting.

Furthermore, when we employ the SwinT pretraining weights [37] on the large-scale ImageNet-1k datasets to initialize our 3DSwinT network, the OA and Kappa of the downstream task can be further improved by 0.03 and 0.05, respectively. To summarize, it can be said that 3DSwinT is more suitable for HSI classification compared to the two SOTA CNN networks. A possible explanation is that 3DSwinT can be boosted with the aid of large-size samples, indicating its great potential for HSI classification.

2) Comparison With SOTA Supervised Methods: All the work in this article is done by pretraining the network by the HCL self-supervised method and then fine-tuning it with a few labels. In this section, we compare the above results with four SOTA supervised methods dedicated to HSI classification, specifically 3-D-FCN [73], SSRN [74], FPGA, and SSDGL. Results show (see Table VI) that HCL self-supervision surpasses the SOTA supervised learning methods and achieves the best classification performance, which can be attributed to the strength of feature learning of the self-supervised pretraining. Consequently, this experiment further demonstrates the superiority and potential of our proposed HCL contrastive learning method.

3) Impact of Data augmentations on Contrastive Learning Performance: Contrastive learning relies heavily on data augmentations since they can provide the essential labels for contrast. We investigate the importance of the two kinds of basic data augmentations used in this study: geometric transformation (including random cropping and scaling, flipping, and rotation) and color space augmentation (including color distortion, random blurring, and graying). As shown in Table VII, geometric and color augmentations play important roles, and both are of almost equal importance. The former enables the network to learn spatially invariant features, while the latter can be used to simulate temporal change features of ground objects, which are two significant characteristics of RS images. Therefore, data augmentation for contrastive learning needs to balance both spatial and color transformations.

Furthermore, we quantitatively discuss and analyze the effects of the three geometric transformation operations, i.e., rotation, crop (random crop and resize), and flip, on the classification results (see Table VIII). Experiments show that, in general, all the geometric transformations can boost the performance of contrastive learning with the increment of OA ranging from 0.4% to 1.1%. The random crop and resize achieve the best result, followed by the flip. Results demonstrate that the geometric transformations based on the whole image are effective for contrastive learning. Geometric augmentations can make the contrastive prediction task more informative and, hence, guide the network to learn high-quality representations.

V. Conclusion

In this article, we proposed a 3DSwinT-HCL method for HSI classification. The proposed 3DSwinT backbone considers the 3-D properties of HSI and can extract rich spatial and spectral features. The HCL framework can adapt to the complex and variable multiscale features of ground objects and adaptively mine multilevel semantic information from unlabeled data. In addition, the MS-LCL module can learn pixel-level information for the dense prediction task. Our research also showed that self-supervised fine-tuning can achieve significantly better accuracy than supervised learning with a small number of labels. Moreover, the 3DSwinT-HCL pretrained model can be well transferred to other hyperspectral datasets, and classification performance for all datasets was improved to some extent.

This study also has limitations. It should be admitted that, in our experiments, the data used for SSL was not very large, owing to the difficulty in collecting large-scale HSIs with sufficient and dense semantic labels. However, our results showed that the self-supervised model could be more robust and sophisticated when more pretraining data were available. In the future, we plan to perform the SSL with more and broader data and transfer the pretrained network to semantic segmentation tasks with different image landscapes. We shall also consider applying the pretrained 3DSwinT-HCL model to more downstream tasks, such as change detection, instance, and segmentation.

Acknowledgment

The authors would also like to thank the editors and anonymous reviewers for the insightful suggestions, which significantly improved the quality of this article.

References

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. A. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2012.

[2] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2017, doi: 10.1109/TGRS.2017.2765364.

[3] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4117–4128, Sep. 2016, doi: 10.1109/JSTARS.2016.2577339.

[4] W. Li and Q. Du, "Joint within-class collaborative representation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2200–2208, Jun. 2014, doi: 10.1109/JSTARS.2014.2306956.

[5] 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, doi: 10.1109/TGRS.2004.831865.

[6] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[7] J. Li, J. Bioucas-Dias, and A. Plaza, "Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[8] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 192–195, Apr. 2005.

[9] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018, doi: 10.1109/TGRS.2018.2828161.

[10] Y. Zhang, X. Wang, X. Jiang, and Y. Zhou, "Marginalized graph self-representation for unsupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2021.3121671.

[11] S. Li and H. Qi, "Sparse representation based band selection for hyperspectral images," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2693–2696.

[12] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[13] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.

[14] X. Guo, X. Huang, L. Zhang, L. Zhang, A. Plaza, and J. A. Benediktsson, "Support tensor machines for classification of hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3248–3264, Jun. 2016.

[15] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.

[16] X. Zhang, X. Jiang, J. Jiang, Y. Zhang, X. Liu, and Z. Cai, "Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022, doi: 10.1109/TGRS.2021.3057701.

[17] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[18] J. Fan, T. Chen, and S. Lu, "Superpixel guided deep-sparse-representation learning for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3163–3173, Nov. 2018, doi: 10.1109/TCSVT.2017.2746684.

[19] 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.

[20] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019, doi: 10.1109/TGRS.2019.2899129.

[21] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[22] M. E. Paoletti *et al.*, "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019, doi: 10.1109/TGRS.2018.2871782.

[23] 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, doi: 10.1109/TGRS.2020.3015157.

[24] V. Slavkovikj, S. Verstockt, W. De Neve, S. Van Hoecke, and R. Van De Walle, "Hyperspectral image classification with convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1159–1162.

[25] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[26] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[27] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.

[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[29] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[30] M. Chen *et al.*, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[31] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[33] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Sep. 2019.

[34] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[35] D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[36] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[37] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[39] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, May 2019.

[40] 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.

[41] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[42] Y. Chen and L. Bruzzone, "Self-supervised change detection by fusing SAR and optical multi-temporal images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3101–3104.

[43] H. Li *et al.*, "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[44] X. Li, D. Shi, X. Diao, and H. Xu, "SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[45] H. Xu, W. He, L. Zhang, and H. Zhang, "Unsupervised spectral–spatial semantic feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2022.3159789.

[46] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9640–9649.

[47] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[48] Z. Xie *et al.*, "Self-supervised learning with Swin transformers," 2021, *arXiv:2105.04553*.

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

[50] 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.

[51] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[52] 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.

[53] 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.

[54] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[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] 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.

[57] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sens.*, vol. 11, no. 2, p. 159, 2019.

[58] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[59] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2020.

[60] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[61] X. Chu *et al.*, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.

[62] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[63] X. Zhao, J. Guo, Y. Zhang, and Y. Wu, "Memory-augmented transformer for remote sensing image semantic segmentation," *Remote Sens.*, vol. 13, no. 22, p. 4518, Nov. 2021.

[64] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, 2021.

[65] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12009–12019.

[66] W. Wang *et al.*, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, pp. 415–424, Mar. 2022.

[67] H. Cao *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[68] L. Gao *et al.*, "STransFuse: Fusing Swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[69] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, p. 3585, Sep. 2021.

[70] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1993, pp. 1–8.

[71] S. Atito, M. Awais, and J. Kittler, "SiT: Self-supervised vIsion transformer," 2021, *arXiv:2104.03602*.

[72] Q. Zhu *et al.*, "A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Trans. Cybern.*, early access, May 25, 2021, doi: 10.1109/TCYB.2021.3070577.

[73] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[74] 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.

Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009.

He is currently a Full Professor with Wuhan University, where he teaches remote sensing, image interpretation, and so on. He is the Head of the School of Remote Sensing and Information Engineering, Institute of Remote Sensing Information Processing (IRSIP), Wuhan University. He was supported by the National Program for Support of Top-Notch Young Professionals in 2017, the China National Science Fund for Excellent Young Scholars in 2015, and the New Century Excellent Talents in University from the Ministry of Education of China in 2011. He has published more than 200 peer-reviewed articles (Science Citation Index (SCI) papers) in international journals. His research interests include remote sensing image processing methods and applications.

Dr. Huang has been an Editorial Board Member of the *Remote Sensing of Environment* since 2019. He was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the Winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was a Lead Guest Editor of the special issue for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *Journal of Applied Remote Sensing*, *Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing*. He was an Associate Editor of the *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2014 to 2020, and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2018 to 2022. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2022.

Mengjie Dong received the B.S. degree from the Taiyuan University of Technology, Taiyuan, China, in 2020. She is currently pursuing the M.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

Her research interests include hyperspectral imagery, land cover classification, semantic segmentation, and deep learning.

Jiayi Li (Senior Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 60 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision, pattern recognition, and remote sensing images.

Dr. Li is the Young Editorial Board Member of *Geo-Spatial Information Science* (GSIS), a Guest Editor of the *Remote Sensing* (an open-access journal from MDPI), and *Sustainability* (an open-access journal from MDPI). She is also a Reviewer for more than 30 international journals, including IEEE Transactions on Geoscience and Remote Sensing (TGRS), IEEE Transactions on Image Processing (TIP), IEEE Transactions on Cybernetics (TCYB), Remote Sensing of Environment (RSE), and ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS-J).

Xian Guo received the B.S. degree in surveying and mapping from Central South University, Changsha, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2015.

He is currently a Lecturer with the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China. His research interests include urban remote sensing, high-resolution image processing, and digital twins modeling.