HSIMAE: A Unified Masked Autoencoder With Large-Scale Pretraining for Hyperspectral Image Classification

Yue Wang, Ming Wen, Hailiang Zhang, Jinyu Sun, Qiong Yang, Zhimin Zhang, and Hongmei Lu

***Abstract*—With a spurt of progress in deep learning techniques, convolutional neural network-based and transformer-based methods have yielded impressive performance on the hyperspectral image (HSI) classification tasks. However, pixel-level manual annotation is time-consuming and laborious, and the small amount of labeled HSI data brings challenges to deep learning methods. Existing methods use carefully designed network architectures combined with self-supervised or semi-supervised learning to deal with the lack of training samples. Those methods were designed for specific datasets and often needed to tune hyperparameters on new datasets carefully. To tackle this problem, a unified HSI masked autoencoder framework was proposed for HSI classification. Different from existing works, the hyperspectral image masked autoencoder (HSIMAE) framework was pretrained on a large-scale unlabeled HSI dataset, named HSIHybrid, which contained a large amount of HSI data acquired by different sensors. First, to handle the different spectral ranges of HSIs, a group-wise PCA was applied to extract features of HSI spectra and transform them into fixed-length vectors. Then, a modified masked autoencoder was proposed for large-scale pretraining. It utilized separate spatial–spectral encoders followed by fusion blocks to learn spatial correlation and spectral correlation of HSI data. Finally, to leverage the unlabeled data of the target dataset, a dual-branch finetuning framework that used an extra unlabeled branch for mask modeling learning was introduced. Extensive experiments were conducted on four HSI datasets from different hyperspectral sensors. The results demonstrate the superiority of the proposed HSIMAE framework over the state-of-the-art methods, even with very few training samples.**

***Index Terms*—Hyperspectral image (HSI) classification, large-scale pretraining, masked autoencoder, self-supervised learning, transformer.**

I. Introduction

HYPERSPECTRAL imaging is a rapid development imaging spectroscopic technique in recent years. It combines traditional imaging technology and spectroscopy to acquire both spatial and spectral information of a target area simultaneously [1]. With high spectral resolution, hyperspectral imaging can capture hundreds of narrow and continuous spectral bands, including visible, near-infrared, and mid-infrared parts. The addition of spectral information is conducive to the fine distinction between different land-cover objects, which is difficult to distinguish in traditional optical images. To this end, hyperspectral imaging has been used in a wide range of fields, such as mineralogy, geology, environment, agriculture, biomedical imaging, and archaeology [2], [3], [4], [5], [6].

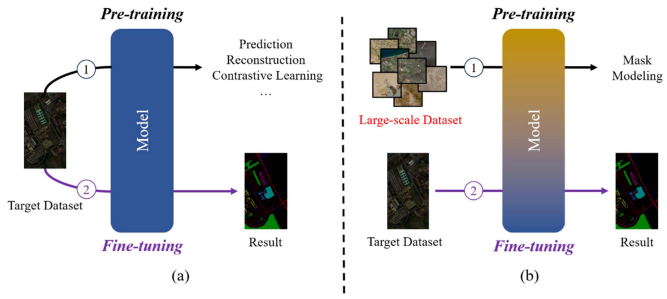

Fig. 1. Difference between HSIMAE and existing self-supervised methods in HSI classification. (a) Most of the existing methods are designed for pretraining and finetuning the classification model on the same dataset, and the model needed to be pretrained from scratch on a new dataset. Differently, (b) HSIMAE is pre-trained by a diverse and large-scale HSI dataset and then directly finetuned on the target dataset.

As a fundamental task of hyperspectral imaging analysis, hyperspectral image (HSI) classification aims to assign a predefined category to each pixel in HSIs. Since HSIs often contain thousands or even millions of pixels, pixel-level manual annotation is time-consuming and laborious. Thus, researchers turned to efficient computer-aided HSI classification. At an early age, support vector machines (SVM) [7] and random forests [8] were often used for HSI classification [9], [10]. Recently, deep learning methods have made breakthroughs in HSI classification [11], whereas the small amount of labeled HSI data brings significant challenges to those methods. A promising solution for this challenge is to leverage the unlabeled data to assist HSI classification. Current research mainly focuses on using carefully designed network architectures combined with self-supervised learning [12], [13], [14] or semi-supervised learning [15], [16] to deal with the lack of training samples. However, many of them were designed for specific datasets and needed to tune hyperparameters on new datasets carefully (see Fig. 1). This

Manuscript received 29 May 2024; revised 9 July 2024; accepted 20 July 2024. Date of publication 23 July 2024; date of current version 15 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFF0906702, and in part by the National Natural Science Foundation of China under Grant 22273120. *(Corresponding authors: Zhimin Zhang; Hongmei Lu.)*

The authors are with the College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China (e-mail: zmzhang@csu.edu.cn; hongmeilu@csu.edu.cn).

The source code is available at https://github.com/Ryan21wy/HSIMAE.

Digital Object Identifier 10.1109/JSTARS.2024.3432743

raises a new question: *How to train a unified HSI classification model from a large amount of unlabeled data*?

As a part of self-supervised representation learning, mask modeling methods have made inspiring progress in various domains [17], [18], [19], [20], [21]. Mask modeling allows models to learn the relationship between local patterns by reconstructing the masked patches/tokens based on the unmasked part. In this study, masked autoencoder (MAE) [20] was chosen as the base architecture for large-scale pretraining. MAE uses an asymmetric autoencoder architecture with a heavy encoder and a lightweight decoder. In the pretraining phase, only the unmasked patches are passed through the encoder. Then, the lightweight decoder completes the reconstructed process. By masking a large proportion of patches (e.g., 75%), MAE can drastically reduce memory consumption, making it easy to scale to large models. In the HSI classification domain, many researchers have applied MAE to improve classification accuracy [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. However, these methods often used the same dataset for pretraining and finetuning, and few of them exploited the scalability of MAE. In this study, we want to explore the potential of MAE by performing large-scale pretraining on the HSI classification domain.

To perform large-scale pretraining, we collected a large-scale HSI dataset named HSIHybrid, which consisted of 15 HSI datasets from different hyperspectral sensors. All of them were freely downloaded from the Internet. As HSI classification models used small data patches as input, the data in the HSIHybrid dataset were segmented into patches with a spatial size of 9×9. A total of 4 million HSI patches were obtained for pretraining. This large and diverse dataset enabled MAE to achieve better and more robust downstream dataset performance.

Large-scale pre-training on HSI data is particularly challenging due to the different spectral resolutions and spectral ranges between different hyperspectral sensors. In order to handle this issue, a group-wise principal component analysis (PCA) was used to transform the raw spectra to fixed-length features, and the mask modeling process was performed on PCA features. Precisely, HSI was first divided equally into different numbers of groups along the spectral dimension. Then, PCA was separately applied to each group. After applying group-wise PCA, the HSI data of different spectral bands were transformed to a uniform size, and the redundant information of hyperspectral data was reduced. Moreover, as the retained principal components are much smaller than the number of spectral bands, the memory occupation was drastically reduced.

Existing MAEs usually use the standard transformer encoder as the backbone network [18], [20], [21]. For HSI classification, MAEST [22] performed pixel-level/patch-level random masking directly on the raw HSI data, which treated tokens at each location equally. FactoFormer [31] extended MAE by using two separate spectral and spatial transformers and performed masked modeling pretraining on spatial dimension and spatial dimension, respectively. This modification allows the model to learn different properties of spectral dimension and spatial dimension. However, FactoFormer required pretraining the spectral and spatial transformer, respectively, increasing the complexity of the method. In this study, a modified transformer encoder was proposed. It consisted of separate spatial-spectral encoders (SSSE) and fusion blocks that simultaneously learned the spatial correlation and spectral correlation of HSI data in an end-to-end manner. Based on the modified MAE encoder, a new MAE variant named hyperspectral image masked autoencoder (HSIMAE) was built. HSIMAE consists of pure transformer blocks, where the feed-forward network (FFN) layer was replaced by the SwiGLU [32] layer. For mask modeling pretraining, a spatial–spectral masking strategy was designed to adapt the separate spatial–spectral encoder. It used successive spatial and spectral masking to ensure the spatial consistency and spectral consistency of the unmasked part. Then, a lightweight decoder received the encoder features and mask token to complete the reconstruction process.

To better utilize the unlabeled data in the target dataset, a dual-branch (a labeled branch and an unlabeled branch) finetuning framework was proposed. For the labeled branch, the decoder of MAE was discarded, and a classification head was added. The labeled data were passed through the encoder and then fed into the classification head for supervised classification. For the unlabeled branch, the unlabeled HSI data completed a full MAE process for self-supervised reconstruction. The encoder was shared between two branches. When training with a small amount of labeled data, the unlabeled branch was used as regularization to suppress overfitting.

HSIMAE, an extension of MAE in the field of hyperspectral imaging, was proposed to train a unified HSI classification model from a large amount of unlabeled data. The main contributions of this work are as follows:

1) A large and diverse HSI dataset named HSIHybrid was curated for large-scale HSI pretraining. It consisted of 15 HSI datasets from different hyperspectral sensors. Experiments showed that increasing the diversity of the pretraining datasets could improve the downstream classification performance.
2) Separate spatial–spectral encoders were designed to learn the spatial correlation and spectral correlation of HSI data. Subsequently, a modified MAE named HSIMAE that utilized separate spatial–spectral encoders followed by fusion blocks was proposed. To handle the different spectral resolutions and spectral ranges of the HSIHybrid dataset, a group-wise PCA was used to extract features of HSI spectra and transform the raw spectra to fixed-length features. Thus, HSIMAE could be pretrained on various HSI datasets with different spectral resolutions and spectral ranges.
3) A dual-branch finetuning framework was introduced to leverage the unlabeled data of the downstream HSI dataset and suppress overfitting on small training samples. The unlabeled branch further adapted HSIMAE to the data distribution of the target dataset. Moreover, the dual-branch framework can be easily adapted to other mask pretraining models to improve performance.
4) Extensive experiments verified the effectiveness of the core components of the HSIMAE framework. With minimal hyperparameter tuning, HSIMAE outperformed the state-of-the-art methods on four public hyperspectral

datasets, even with very few training samples. This demonstrated that the HSIMAE framework was a promising solution to build a unified HSI classification model.

The rest of this article is organized as follows. In Section II, some related works are shown. HSIMAE is described in detail in Section III. The experiments and results are presented in Section IV followed by the discussion in Section V. Finally, Section VI concludes this article.

II. Related Work

A. Deep Learning in HSI Classification

Recently, with its powerful representation learning ability, deep neural networks (DNNs) have been widely applied to hyperspectral classification tasks. According to the characteristics of HSI, deep learning-based classification methods can be divided into spectral-based, spatial-based, and joint spatial–spectral methods. Spectral-based methods followed the simple intuition of directly classifying pixels based on their spectra. Those methods extracted the pixel vector of the HSI by a one-dimensional (1-D) convolutional neural network (CNN) [33], [34], 1-D generative adversarial network [35], or recurrent neural network [36], [37]. In contrast, spatial-based methods employed an "image-like" classification strategy to process HSI. The 2-D representation of HSI was extracted by PCA. Subsequently, 2-D CNN was used to classify the target pixel according to its neighborhood region [34], [38]. Using off-the-shelf models pretrained on the ImageNet [39] as initialization was an effective strategy to improve the accuracy of the spatial-based methods [40]. Joint spatial–spectral methods integrated spectral and spatial information from HSIs to improve classification accuracy, surpassing approaches that focus on a single aspect. The spectral and spatial features were extracted by 3D-DNN [34], [41], [42] or by two separate 1D-DNN and 2D-DNN [43], [44], then fused for classification.

As great successes in natural language processing, transformer-based methods have gained tremendous attention in recent years [45], [46]. Transformers can obtain global representations through their long-distance modeling capabilities, which are considered suitable for HSI classification tasks [47], [48], [49], [50], [51], [52]. As the first attempt, HSI-BERT [47] treated pixel spectra in HSI as "words" in language and used BERT [17] to capture the global dependence among spectra. Instead, SpectralFormer [48] introduced vision transformer [46] to learn group-wise spectral embeddings by grouping HSI in spectral dimension and using long-term skip connections to learn cross-layer feature fusion adaptively. Due to the lack of inductive bias and small labeled samples, pure transformers performed worse than their CNN counterparts. Recent works focused on combining the advantages of CNN and transformer. Some of them used CNN to extract local features followed by transformer blocks to model global relationships [53], [54], [55], [56], [57], [58]. Some modified the transformer blocks by adding convolution modules to realize the extraction of local and global information [59], [60], [61], [62]. The others used separate CNN branch and transformer branch to extract local and global features, separately, and then fused the features for HSI classification [63], [64].

B. Self-Supervised Learning in HSI Classification

Since obtaining fine-grained pixel-wise annotations is expensive and laborious, self-supervised learning methods are emerging in HSI classification [14]. It aims to extract robust features from unlabeled data by carefully designing pretext tasks. According to the pretext task, self-supervised learning can be mainly divided into two categories: generative learning and contrastive learning. Generative learning methods learn feature representations by reconstructing the original inputs from corrupted ones, including autoencoders [12], [65], [66] and generative adversarial networks [13]. Contrastive learning methods learn invariant semantic representation by aligning features from different views of the same input [50], [67], [68], [69], [70], [71].

Recently, as a kind of generative learning, mask autoencoders have received tremendous interest in hyperspectral classification [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. MAEST [22] performed pixel-level/patch-level masking directly on the raw HSI data and used the SpectralFormer architecture for classification. SSLSM [26] first applied PCA to reduce the spatial dimension and then used spectral masking for pretraining. IMAE [30] used spatial–spectral embedding to transform raw HSI data and added position embedding by conditional position embedding. Moreover, it selected HSIs gathered by the GaoFen-5 satellite as the pre-training dataset, and a total of 300 000 samples were obtained by dividing the HSI data into different sizes. FactoFormer [31] used two separate spectral and spatial Transformers and performed masked modeling pretraining on spatial dimension and spatial dimension, respectively. In the finetuning phase, the outputs of the spectral and spatial transformers were pooled and concatenated as the classification features.

In this work, we focus on exploring a unified mask autoencoder framework with large-scale pretraining for HSI classification. Although IMAE was pretrained on an HSI dataset with a total of 300 000 samples, the size of the pretraining dataset was still too small compared to the image and video domain. Moreover, the pretraining dataset only contained HSIs gathered by the GaoFen-5 satellite, limiting the generalization ability of IMAE. In contrast, we collected a large-scale and diverse HSI dataset named HSIHybrid, consisting of 15 HSI datasets from different hyperspectral sensors and containing a total of 4 million HSI patches for pretraining. The substantially increased size and diversity of the HSIHybrid dataset significantly contributed to the enhanced performance and robustness of HSIMAE across various downstream applications. Instead of learning spatial and spectral features separately in FactoFormer, we explored a modified MAE architecture to learn both spatial correlation and spectral correlation simultaneously and using additional fusion blocks to fuse the spatial and spectral features further.

III. Methodology

The proposed HSIMAE is an extension of MAE in the hyperspectral imaging field. It is an asymmetric encoder–decoder network that reconstructs the masked data from the observation data. Given an HSI as input, HSIMAE first used group-wise PCA to transform the raw spectra into fixed-length features,

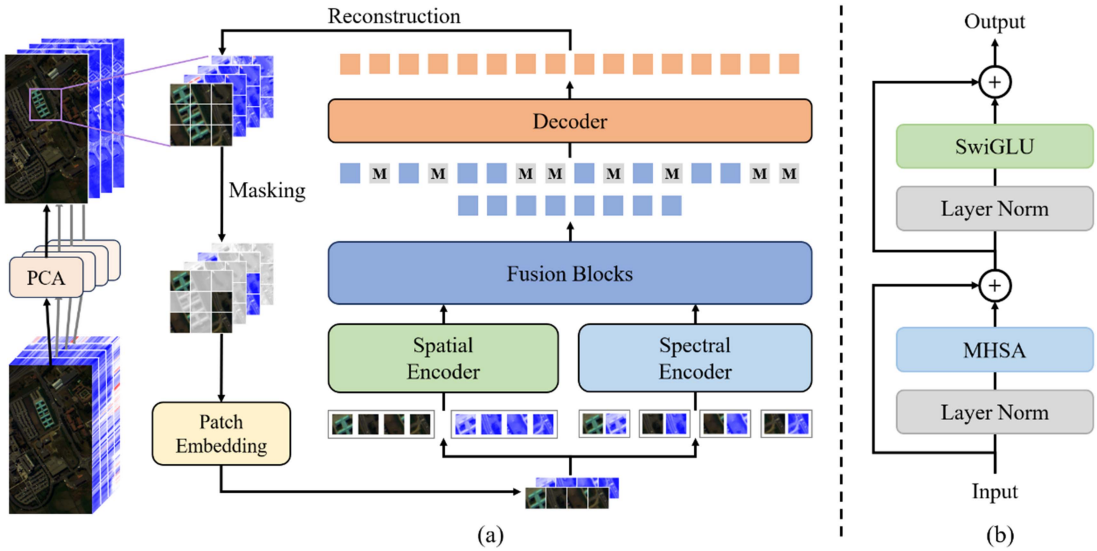

Fig. 2. Overview of the proposed HSIMAE pre-training framework. (a) HSI is first transformed into fixed-length features by group-wise PCA, and the features are divided into small data patches. Then, a 3-D cube embedding layer partitions the data patch into nonoverlapping cubes and masks a subset of them. The unmasked cubes are divided into groups based on spatial position and spectral position, and fed to the spatial and spectral encoder, respectively. The spatial and spectral features are further fused by the fusion blocks. After that, mask tokens are added to the masked position. A small decoder then processes both the patch features and mask tokens to reconstruct the PCA features. (b) Encoder and decoder of HSIMAE both use the vanilla ViT block, which contains multihead self-attention, SwiGLU, and layer normalization.

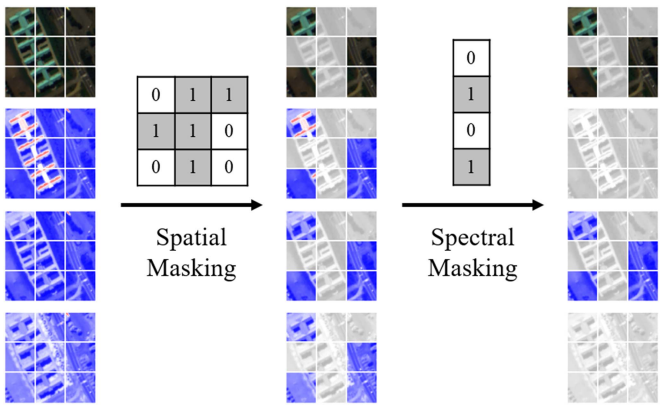

Fig. 3. Spatial-spectral masking strategy. For a given 3D cube, spatial masking first masks out all cubes at the same spatial location. Then, spectral masking masks out all tokens at the same spectral location. Thus, spectral consistency and spatial consistency are guaranteed during the masking process.

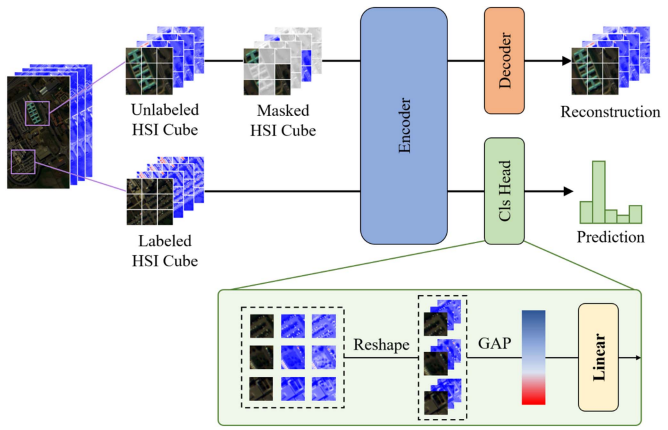

Fig. 4. Overview of the introduced dual-branch finetuning framework. For the labeled branch, the encoder of HSIMAE is reserved for feature extraction, and an extra classification head is added for the HSI classification task. The feature tokens of the labeled cubes at the same spatial location are concatenated. Global average pooling (GAP) is then applied to the concatenated tokens to obtain the global representation vector. For the unlabeled branch, the unlabeled data go through the entire MAE process. Two branches are joint learning during the finetuning process.

then partitioned these features into consecutive 3-D patches and masked part of them. The encoder of HSIMAE mapped the unmasked parts to feature vectors, and the lightweight decoder reconstructed the masked parts from these feature vectors. In the finetuning phase, a dual-branch finetuning framework was used for HSI classification with small samples. For the labeled branch, the HSIMAE decoder was discarded, and an extra classification head was equipped. The unlabeled branch followed the MAE architecture. The encoder parameters were shared between two branches. The following section describes three core modules of HSIMAE in detail: the pretraining framework (see Fig. 2), spatial–spectral masking (see Fig. 3), and dual-branch finetuning framework (see Fig. 4).

A. *Hyperspectral Imaging Masked Autoencoder Pretraining Framework*

The different spectral resolutions and spectral ranges of HSI datasets made large-scale pretraining on HSI data challenging. A group-wise PCA was used to handle this issue. Given an HSI $I \in R^{H\times W\times B}$, it was divided equally into T groups in the spectral dimension. Then, PCA extracted features of length L for each group, respectively. Finally, the features were concatenated as the output $I_P \in R^{H\times W\times T\cdot L}$. The group-wise PCA process can

be formulated as follows:

$$\begin{aligned} I_1, I_2, \ldots, I_T &= \text{Split}(I), \\ P_i &= \text{PCA}(I_i), \forall i \in [1, T], \\ I_P &= \text{Concatenate}(P_1, P_2, \ldots, P_T). \end{aligned} \tag{1}$$

In this study, T was 4, and L was 8. The length of the final PCA features was 32.

After group-wise PCA processing, HSI was transformed into fixed-length features $I_P \in R^{H\times W\times T\cdot L}$. Since HSIs were usually large in size and rich in local information, the PCA features were first divided into small data patches $I_p \in R^{h\times w\times T\cdot L}$ along the spatial dimension. Following the idea of VideoMAE [21], a 3-D cube embedding was adopted in HSIMAE, where each 3-D cube of size $s \times s \times L$ was treated as one token embedding. The data patch was partitioned into nonoverlapping cubes $x \in R^{\frac{h}{s}\times\frac{w}{s}\times T\times s^2 L}$. Then, a part of the cubes was masked using a spatial–spectral masking strategy. Subsequently, a trainable linear projection mapped each cube into a C-dimensional token, and a separable sinusoidal positional embedding [45] $E_{\text{pos}} \in R^{\frac{h}{s}\times\frac{w}{s}\times T\times C}$ was added to all tokens, which was a combination of 1-D sinusoidal spectral positional embedding and 2-D sinusoidal spatial positional embedding. The positional embedding enabled the model to perceive the location of each token. It was defined by the following equation:

$$\begin{aligned} x_{um}, x_m &= \text{Masking}(x), \\ z_{um}, z_m &= \text{Linear}([x_{um}, x_m]) + E_{\text{pos}}. \end{aligned} \tag{2}$$

The masked tokens were discarded, and only the unmasked tokens were fed into the encoder. The unmasked tokens can be represented as $z_{um} \in R^{l\times t\times C}$, where $l = \frac{h}{s} \times \frac{w}{s} \times (1 - mr_{\text{spa}})$, $t = T \times (1 - mr_{\text{spe}})$, mr_{spa}, and mr_{spe} denoted the spatial and spectral mask ratios, respectively. The unmasked tokens were divided into t spatial groups and l spectral groups, then fed to the spatial and spectral encoder, respectively. The spatial and spectral features were further fused by the fusion blocks. This process can be represented as follows:

$$\begin{aligned} z^1_{\text{spa}}, z^2_{\text{spa}}, \ldots, z^t_{\text{spa}} &= \text{SpatialGroup}(z_{um}), \\ z^1_{\text{spe}}, z^2_{\text{spe}}, \ldots, z^l_{\text{spe}} &= \text{SpectralGroup}(z_{um}), \\ {z^i_{\text{spa}}}' &= \text{SpatialEncoder}(z^i_{\text{spa}}), \forall i \in [1, t], \\ {z^j_{\text{spe}}}' &= \text{SpectralEncoder}(z^j_{\text{spe}}), \forall j \in [1, l], \\ {z_{um}}' &= \text{FusionBlock}({z_{\text{spa}}}' + {z_{\text{spe}}}'). \end{aligned} \tag{3}$$

Then, a linear layer projected the feature tokens of the unmasked tokens to the decoder dimension C_D. Their average pooled token was used as the mask token $M \in R^{C_D}$ and filled the position of the masked positions. An extra separable sinusoidal positional embedding $E^D_{\text{pos}} \in R^{\frac{h}{s}\times\frac{w}{s}\times T\times C_D}$ was added. Finally, the lightweight decoder reconstructed the mask part of the data patch. The process can be represented by the following expression:

$$\begin{aligned} z^D_{um} &= \text{Linear}({z_{um}}'), \\ [\hat{x}_{um}, \hat{x}_m] &= \text{Decoder}([z^D_{um}, M] + E^D_{\text{pos}}) \end{aligned} \tag{4}$$

where $\hat{x}_{um}$ and $\hat{x}_m$ are the reconstructed cubes of unmasked patch tokens and masked patch tokens, respectively.

The target of the masked HSI modeling pre-training is to reconstruct the masked part of the raw data. Following MAE, the mean squared error (MSE) between the reconstructed and original cubes with per-patch normalization was used as the loss function. Only the mask part was used to calculate the loss. The MSE loss can be formulated as follows:

$$L_{\text{rec}} = \frac{1}{M} \sum_{i\in M} \|x_{m,i} - \hat{x}_{m,i}\|_2^2 \tag{5}$$

where M is the number of masked cubes.

As shown in Fig. 2(b), a modified ViT block [46] was used as the encoder and decoder of HSIMAE. Specifically, the FFN layer was replaced by SwiGLU [32] layer, which can be formulated as follows:

$$\begin{aligned} \text{Swish}_\beta(z) &= z * \text{Sigmoid}(\beta z), \\ \text{SwiGLU}(z) &= \text{Swish}_1(zW + b) \otimes (zV + c) \end{aligned} \tag{6}$$

where W and V are the parameter matrices, and b and c are bias vectors.

The block used in HSIMAE can be formulated as follows:

$$\begin{aligned} {z'}_{l-1} &= \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \forall l \in [1, L], \\ z_l &= \text{SwiGLU}(\text{LN}(z'_{l-1})) + z'_{l-1}, \forall l \in [1, L] \end{aligned} \tag{7}$$

where L is the number of blocks and $z_l \in R^{N\times C}$ is the output tokens of the l block, and LN represents layer normalization.

B. Spatial–Spectral Masking

The separate spatial and spectral encoders in HISMAE required spatial consistency and spectral consistency of the input data, which means that the spatial locations of the unmasked cubes were the same at all spectral locations and vice versa. Random masking did not guarantee this consistency. Thus, a simple spatial–spectral masking strategy was proposed. It consisted of two successive masking processes: spatial masking and spectral masking. For a given 3-D cube, spatial masking first masked out all cubes at the same spatial location. Then, spectral masking masked out all tokens at the same spectral location. Thus, spectral consistency and spatial consistency were guaranteed during the masking process. The total mask ratio can be obtained as follows:

$$\text{Mask Ratio} = 1 - (1 - mr_{\text{spa}}) \times (1 - mr_{\text{spe}}) \tag{8}$$

where mr_{spa} and mr_{spe} denote the spatial and spectral mask ratios, respectively. In particular, at least two spatial locations and two spectral locations were not masked.

C. Dual-Branch Finetuning Framework

To leverage the unlabeled data of the downstream HSI dataset, a dual-branch finetuning framework was introduced that used an additional unlabeled branch parallel to the labeled classification branch. In the finetuning phase, HSIs were transformed into

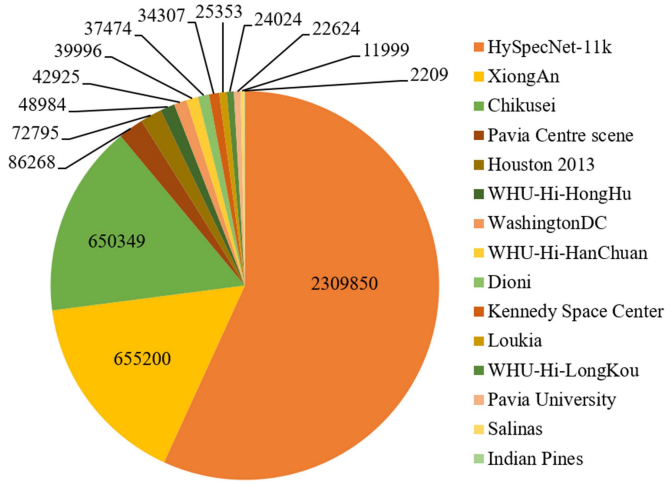

Fig. 5. Data distribution of the HSIHybrid dataset.

fixed-length features by group-wise PCA and were divided into data patches. The label of each data patch was determined by the center pixel.

For the labeled branch, the encoder of HSIMAE was reserved for feature extraction, and the decoder was discarded. A classification head was used for the HSI classification task. To better utilize the group-wise PCA features, the feature tokens at the same spatial location were concatenated. Global average pooling (GAP) was then applied to the concatenated tokens to obtain the global representation vector. A linear layer was used as a classifier to predict classification results.

For the unlabeled branch, unlabeled data went through the entire MAE process using the same encoder. The unlabeled data were obtained by dividing the group-wise PCA features into nonoverlapping patches. In each mini-batch, half of the total unlabeled data were sampled to calculate the reconstruction loss. In practice, the labeled cubes were also considered as part of unlabeled data. Cross entropy loss was used as the classification loss L_{cls}, and MSE loss as the reconstruction loss L_{rec}. The two branches were joint training at each iteration. The total loss can be formulated as follows:

$$L = L_{\text{cls}} + \lambda L_{\text{rec}} \tag{9}$$

where λ is a fixed scalar hyperparameter denoting the relative weight of the unlabeled reconstruction loss.

IV. Experiments and Results

A. Dataset Description

Here, a large-scale HSI dataset named HSIHybrid was first constructed. As shown in Fig. 5, the HSIHybrid dataset consisted of 15 HSI datasets from different hyperspectral sensors. As HSI classification models used small data patches as input, the data in the HSIHybrid dataset were segmented into patches with a spatial size of 9×9. Specifically, since the HyspecNet-11k dataset [72] was much larger than the other 14 HSI datasets, oversampling was used for the other 14 HSI datasets, which would alleviate overfitting of the model to the distribution of the HyspecNet-11k dataset. As a result, a total of 4 064 357 (4 million) image patches were obtained for HSIMAE pretraining.

Four HSI classification datasets from different hyperspectral sensors were used to evaluate the performance of the proposed HSIMAE, including Salinas, Pavia University, Houston 2013, and WHU-Hi-LongKou datasets. Table I lists the land-cover class names and the number of labeled samples of four classification datasets. The details of each dataset are described as follows.

1) The Salinas dataset was collected by the airborne visible/infrared imaging spectrometer sensor over Salinas Valley, California. The raw imagery consisted of 224 bands in the range of 400 to 2500 nm. The size of the imagery was 512 × 217 pixels with a spatial resolution of 3.7 m. In this study, 20 water absorption bands were discarded, and 204 bands were reserved. It contains 16 land-cover classes, and 54 129 pixels were labeled.
2) The Pavia University dataset was gathered by the reflective optics system imaging spectrometer sensor over the city of Pavia, Italy. The raw imagery consisted of 115 bands in the range of 430 to 860 nm. The size of the imagery was 610 × 340 pixels with a spatial resolution of 1.3 m. In this study, 12 noise bands were discarded, and 103 bands were reserved. The Pavia University dataset contained 9 classes, and 42 776 pixels were labeled.
3) The Houston 2013 dataset was acquired by the NSF-funded Center for Airborne Laser Mapping over the University of Houston campus and the neighboring urban area on 23 June 2012. The hyperspectral imagery consisted of 144 bands in the range of 380 to 1050 nm. The size of the imagery was 349 × 1905 pixels with a spatial resolution of 2.5 m. Fifteen classes were contained, and 15 029 pixels were labeled.
4) The WHU-Hi-LongKou dataset was acquired in Longkou Town, Hubei province, China, with an 8-mm focal length Headwall Nano-Hyperspec imaging sensor equipped on a DJI Matrice 600 Pro (DJI M600 Pro) UAV platform [73]. The size of the imagery was 550 × 400 pixels. There were 270 bands from 400 to 1000 nm, and the spatial resolution of the UAV-borne hyperspectral imagery was about 0.463 m. LongKou dataset contained 9 classes in total, and 204 542 pixels were labeled.

Each dataset was divided into training, validation, and test sets. The training set and validation set contained 20 random samples per class, respectively, and the remaining samples were considered as the test set.

B. Experimental Setup

1) Implementation Details: The proposed HSIMAE framework was implemented on the PyTorch platform. All the experiments were conducted on an Intel Core i9-10900X CPU with 64 GB RAM and an NVIDIA GeForce RTX 3080 10-GB GPU. The vanilla ViT backbone with 12 blocks was adapted as the base architecture of HSIMAE.

The number of blocks in the spectral and spatial encoder was nine, and the remaining three blocks were fusion blocks. A lightweight decoder with 8 blocks and 64 dimensions was

TABLE I
CATEGORY INFORMATION OF FOUR CLASSIFICATION DATASETS

No.	Salinas		Pavia University		Houston 2013		WHU-Hi-LongKou	
	Class	Samples	Class	Samples	Class	Samples	Class	Samples
1	Brocoli green weeds 1	2009	Asphalt	6631	Healthy Grass	1251	Corn	34511
2	Brocoli green weeds 2	3726	Meadows	18649	Stressed Grass	1254	Cotton	8374
3	Fallow	1976	Gravel	2099	Synthetic Grass	697	Sesame	3031
4	Fallow rough plow	1394	Trees	3064	Tree	1244	Broad-leaf soybean	63212
5	Fallow smooth	2678	Painted metal sheet	1345	Soil	1242	Narrow-leaf soybean	4151
6	Stubble	3959	Bare Soil	5029	Water	325	Rice	11854
7	Celery	3579	Bitumen	1330	Residential	1268	Water	67056
8	Grapes untrained	11271	Self-Blocking Bricks	3682	Commercial	1244	Roads and houses	7124
9	Soil vinyard develop	6203	Shadows	947	Road	1252	Mixed weed	5229
10	Corn senesced green weeds	3278			Highway	1227		
11	Lettuce romaine 4wk	1068			Railway	1235		
12	Lettuce romaine 5wk	1927			Parking Lot1	1233		
13	Lettuce romaine 6wk	916			Parking Lot2	469		
14	Lettuce romaine 7wk	1070			Tennis Court	428		
15	Vinyard untrained	7268			Running Track	660		
16	Vinyard vertical trellis	1807						
	Total	54129		42776		15029		204542

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE SALINAS DATASET WITH 20 TRAINING SAMPLES

No.	SVM-RBF	SSRN	FDSSC	DBDA	SpectralFormer	SSFTT	HybridFormer	DCTN	GSC-ViT	HSIMAE-B	HSIMAE-L
1	98.05	100.00	100.00	99.89	92.17	100.00	99.71	100.00	99.81	100.00	100.00
2	96.47	99.99	99.92	99.84	99.25	99.75	100.00	100.00	99.97	99.98	99.96
3	97.19	100.00	99.33	93.49	79.35	100.00	99.99	99.83	99.25	100.00	100.00
4	99.54	99.84	99.45	99.47	97.02	99.79	99.84	99.79	99.75	99.70	99.72
5	96.37	98.62	98.01	94.37	85.80	97.93	98.13	98.86	98.07	98.64	98.51
6	99.06	100.00	99.96	100.00	98.80	99.95	99.75	99.89	99.98	100.00	99.92
7	98.95	99.99	99.89	99.54	96.73	99.85	99.97	99.98	100.00	99.98	99.91
8	70.81	84.71	81.03	73.71	69.55	86.99	89.87	79.30	80.79	90.75	88.92
9	96.95	99.75	99.48	95.73	92.44	100.00	100.00	100.00	99.93	100.00	100.00
10	86.48	96.12	94.72	88.49	87.14	98.13	97.12	96.96	96.57	98.05	98.74
11	95.58	99.90	99.24	99.11	97.82	99.98	100.00	100.00	99.86	100.00	100.00
12	98.57	100.00	99.99	99.84	89.14	99.74	99.95	99.95	99.47	99.94	99.99
13	97.03	99.84	99.84	99.75	93.13	100.00	100.00	100.00	99.86	99.95	99.98
14	94.68	99.13	99.26	99.34	98.10	99.18	99.88	99.81	99.79	99.65	99.75
15	63.09	83.74	87.72	78.75	73.23	95.62	82.42	88.99	92.95	96.04	93.62
16	97.49	98.30	98.23	96.77	95.29	98.92	99.80	99.47	97.92	99.45	99.63
OA (%)	86.60 ± 1.05	94.18 ± 0.77	93.74 ± 1.29	89.70 ± 0.64	85.44 ± 1.37	96.37 ± 1.02	95.18 ± 1.78	93.88 ± 0.84	94.56 ± 0.37	**97.30** ± 0.75	96.62 ±0.96
AA (%)	92.89 ± 0.57	97.50 ± 0.37	97.26 ± 0.70	94.88 ± 0.35	90.31 ± 1.21	98.49 ± 0.39	97.90 ± 0.90	97.68 ± 0.22	97.75 ± 0.25	**98.88** ± 0.31	98.67 ±0.43
K×100	85.10 ± 1.16	93.52 ± 0.86	93.04 ± 1.43	88.56 ± 0.70	83.85 ± 1.54	95.96 ± 1.13	94.63 ± 2.00	93.20 ± 0.93	93.95 ± 0.41	**96.99** ± 0.84	96.24 ±1.07

Best results of the compared methods are underlined. Best results of HSIMAEs are in bold.

equipped. The dimension of each self-attention head was 16 for the encoder and 8 for the decoder. Two HSIMAE models named HSIMAE-B and HSIMAE-L were designed. All the model variants shared the same architecture except for the encoder dimension, which was 128 for HSIMAE-B and 256 for HSIMAE-L, respectively. To improve efficiency and effectiveness, HSIMAE-B pretrained on 10% samples of the HSIHybrid dataset was used for parameter selection, and the determined parameters were applied to HSIMAE-L for method comparison. With regard to group-wise PCA, the number of groups was set to four, and each group had eight8 components. HSIs were transformed into features of length 32 by Group-wise PCA, then cut into patches with the size of 9×9×32. We used AdamW [74] optimizer, cosine learning rate schedule with a warm-up, and drop path ratio of 0.2 in both pretraining and finetuning. In pretraining, the batch size was set to 512, the base learning rate was 0.005, and the weight decay was 0.05. HSIMAE was trained for 10 epochs with a mask ratio of 50%. In fine-tuning, the mask ratio was set to 80%, batch size was 32, the weight decay was 0.005, and the number of epochs was 200. λ was set to 10. The base learning rate was chosen based on the validation results.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE PAVIA UNIVERSITY DATASET WITH 20 TRAINING SAMPLES

No.	SVM-RBF	SSRN	FDSSC	DBDA	SpectralFormer	SSFTT	HybridFormer	DCTN	GSC-ViT	HSIMAE-B	HSIMAE-L
1	76.24	95.74	94.80	93.60	70.43	86.21	93.63	95.41	94.30	95.93	96.51
2	71.11	93.67	92.66	86.55	78.97	91.87	94.76	94.82	93.15	97.07	97.49
3	74.43	93.17	96.27	94.98	66.79	92.74	95.92	94.03	93.64	91.84	95.31
4	89.91	98.13	96.59	95.18	93.64	94.58	95.87	97.64	95.20	97.83	98.06
5	99.25	99.94	99.85	99.80	99.46	99.75	99.68	99.65	99.62	99.77	99.82
6	80.81	99.84	99.34	92.03	70.54	94.42	99.45	99.80	99.14	99.85	99.94
7	89.21	99.61	99.27	98.56	84.62	99.49	99.98	99.95	98.96	99.86	99.80
8	71.29	93.28	95.79	90.92	70.53	83.59	96.24	87.62	94.86	93.98	94.24
9	99.87	99.98	99.63	98.61	99.63	98.37	98.41	99.32	99.40	98.57	98.24
OA (%)	76.60 ± 3.00	95.48 ± 1.62	95.07 ± 2.04	90.72 ± 1.88	77.63 ± 1.52	91.43 ± 2.64	95.79 ± 0.95	95.44 ± 1.77	94.86 ± 1.21	96.95 ± 1.46	**97.44** ±1.61
AA (%)	83.57 ± 1.33	97.04 ± 0.80	97.13 ± 0.77	94.47 ± 0.62	81.62 ± 1.54	93.45 ± 1.49	97.11 ± 0.16	96.47 ± 0.96	96.47 ± 0.71	97.19 ± 0.85	**97.71** ±0.83
K×100	70.39 ± 3.44	94.08 ± 2.09	93.56 ± 2.62	87.95 ± 2.32	71.15 ± 2.04	88.80 ± 3.36	94.46 ± 1.23	94.02 ± 2.28	93.26 ± 1.58	95.98 ±1.90	**96.63** ±2.10

Best results of the compared methods are underlined. Best results of HSIMAEs are in bold.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE HOUSTON 2013 DATASET WITH 20 TRAINING SAMPLES

No.	SVM-RBF	SSRN	FDSSC	DBDA	SpectralFormer	SSFTT	HybridFormer	DCTN	GSC-ViT	HSIMAE-B	HSIMAE-L
1	91.69	95.99	94.53	94.29	94.85	95.24	93.51	95.87	94.81	94.53	95.18
2	97.71	99.21	98.98	95.32	94.38	98.85	98.20	98.60	98.78	98.48	98.06
3	98.51	99.97	99.97	99.85	97.84	99.97	99.76	99.73	99.57	99.82	99.85
4	97.34	99.77	99.65	99.17	95.66	99.22	99.29	98.55	99.45	97.89	98.04
5	94.31	99.95	98.05	96.56	95.22	100.00	99.70	98.84	96.61	100.00	100.00
6	96.84	99.23	98.95	98.39	85.75	98.74	96.98	97.96	98.46	99.09	98.81
7	87.30	93.55	93.60	93.08	75.50	93.50	94.53	96.21	91.19	96.89	97.98
8	65.71	77.51	76.63	75.71	70.81	77.81	82.84	84.02	79.10	84.29	85.78
9	76.85	84.17	85.73	85.79	75.74	86.53	90.99	90.89	87.23	94.16	94.06
10	80.40	88.56	90.67	88.73	81.89	95.74	99.12	99.39	94.76	97.52	97.44
11	83.73	95.63	98.46	90.44	77.92	93.32	98.06	94.59	95.36	96.79	97.62
12	60.37	83.10	84.29	83.52	67.44	88.77	84.96	87.43	84.88	86.19	85.45
13	38.74	94.92	92.96	94.92	72.31	95.57	95.66	95.06	94.41	95.94	95.99
14	97.89	100.00	99.95	99.95	97.78	99.33	99.74	99.64	99.74	100.00	100.00
15	98.65	100.00	100.00	99.84	99.06	100.00	100.00	99.94	99.58	100.00	100.00
OA (%)	84.22 ± 0.93	92.95 ± 2.12	93.14 ± 2.05	91.68 ± 1.01	84.46 ± 1.73	93.89 ± 1.03	94.88 ± 0.68	95.15 ± 0.98	93.26 ± 2.14	95.41 ± 1.08	**95.65** ±0.92
AA (%)	84.40 ± 0.90	94.10 ± 1.69	94.16 ± 1.71	93.04 ± 0.89	85.48 ± 1.50	94.84 ± 0.75	95.56 ± 0.70	95.78 ± 0.84	94.26 ± 1.79	96.11 ± 0.84	**96.28** ±0.71
K×100	82.93 ± 1.00	92.37 ± 2.29	92.59 ± 2.21	91.00 ± 1.09	83.22 ± 1.87	93.40 ± 1.11	94.47 ± 0.73	94.75 ± 1.06	92.71 ± 2.32	95.04 ± 1.17	**95.30** ±1.00

Best results of the compared methods are underlined. Best results of HSIMAEs are in bold.

TABLE V
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE WHU-HI-LONGKOU DATASET WITH 20 TRAINING SAMPLES

No.	SVM-RBF	SSRN	FDSSC	DBDA	SpectralFormer	SSFTT	HybridFormer	DCTN	GSC-ViT	HSIMAE-B	HSIMAE-L
1	89.86	98.50	99.07	98.66	94.97	99.12	96.95	99.53	99.23	99.09	99.68
2	77.14	98.87	98.75	94.75	62.65	99.11	99.33	99.72	99.57	99.66	99.36
3	87.28	99.57	99.63	99.18	90.54	100.00	99.42	99.95	99.30	99.97	100.00
4	76.51	95.39	96.01	93.33	78.56	94.41	90.30	95.92	95.82	96.78	97.89
5	82.94	97.30	98.95	94.26	81.00	98.39	98.24	99.10	97.22	98.96	99.01
6	93.82	99.47	99.12	98.58	95.50	98.60	97.46	98.56	99.22	98.79	99.21
7	99.90	99.71	99.68	99.53	99.61	99.02	99.86	99.50	99.63	98.72	98.73
8	81.69	96.30	95.70	92.41	88.78	89.27	89.94	95.66	95.09	91.87	92.04
9	71.82	96.94	96.08	94.36	83.18	95.04	97.91	98.05	96.38	96.63	96.05
OA (%)	87.81 ± 1.90	97.88 ± 0.74	98.13 ± 0.32	96.73 ± 0.71	89.27 ± 0.88	97.15 ± 0.77	95.81 ± 2.66	98.18 ± 0.36	98.06 ± 0.49	97.96 ± 0.63	**98.41** ±0.51
AA (%)	84.55 ± 1.03	98.01 ± 0.33	98.11 ± 0.43	96.12 ± 0.25	86.09 ± 1.56	97.00 ± 0.29	96.60 ± 1.62	98.44 ± 0.24	97.94 ± 0.25	97.83 ± 0.34	**98.00** ±0.32
K×100	84.39 ± 2.34	97.23 ± 0.95	97.55 ± 0.42	95.73 ± 0.92	86.23 ± 1.11	96.28 ± 0.98	94.58 ± 3.41	97.62 ± 0.46	97.47 ± 0.63	97.33 ± 0.81	**97.92** ±0.67

Best results of the compared methods are underlined. Best results of HSIMAEs are in bold.

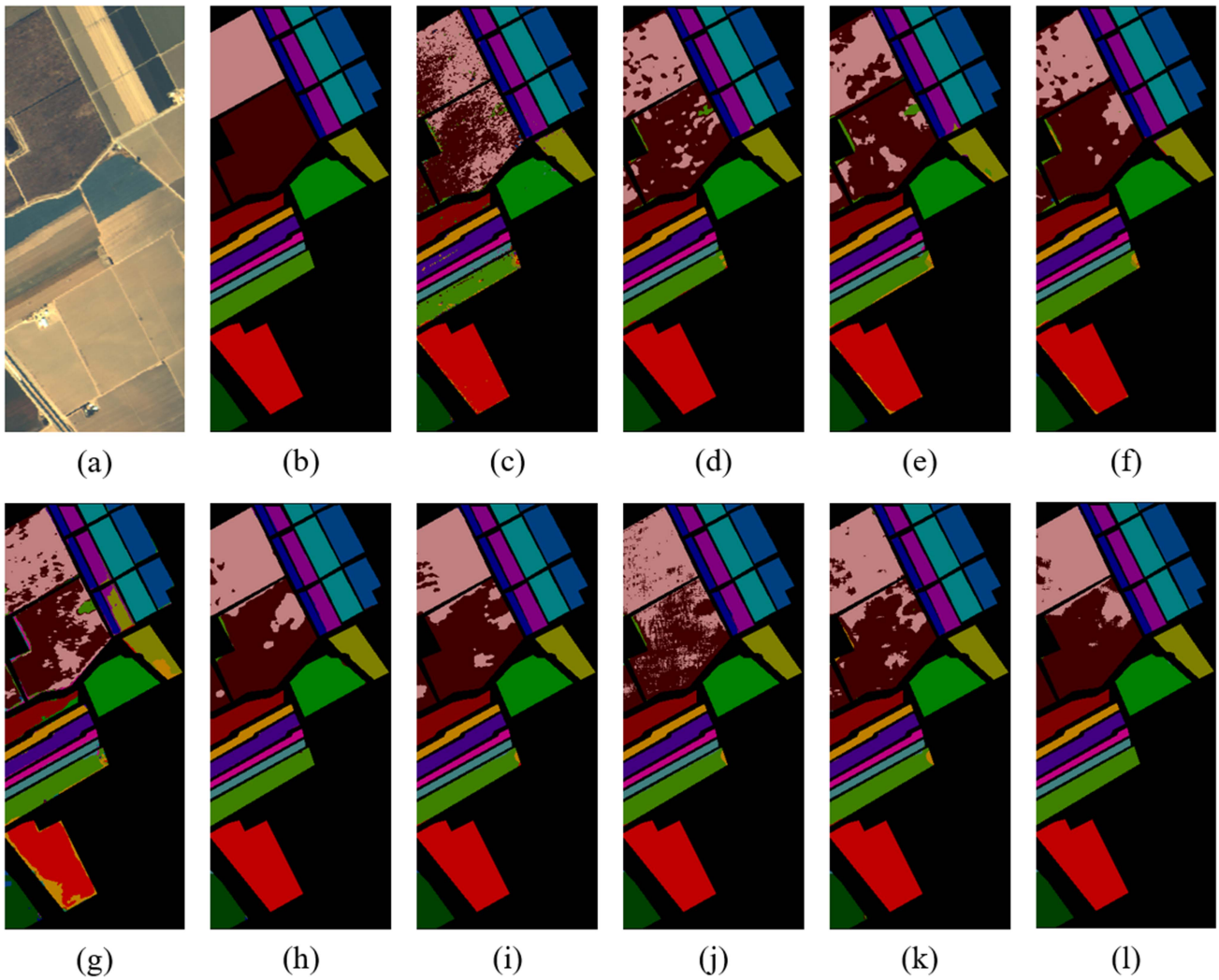

Fig. 6. Classification maps of the full Salinas dataset. (a) Pseudocolor image. (b) Ground truth label. (c) SVM. (d) SSRN. (e) FDSSC. (f) DBDA. (g) SpectralFormer. (h) SSFTT. (i) HybridFormer. (j) DCTN. (k) GSC-ViT. (l) HSIMAE-L.

2) Evaluation Metrics: Four widely used quantitative metrics were adopted in this study, including the accuracy of each class, overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ). To reduce the impact of random initialization, each experiment was conducted 5 times, and the mean and variance of the test results were reported.

3) Compared Methods: To demonstrate the innovation and effectiveness, HSIMAE was compared with nine HSI classification methods. Those methods included one traditional method: SVM-RBF, three CNN-based methods: SSRN [42], FDSSC [75], and DBDA [76], and five transformer-based methods: SpectralFormer [48], SSFTT [53], HybridFormer [55], DCTN [64], and GSC-ViT [62]. For SVM-RBF, the regularization parameter and the kernel parameter were determined by a grid search using the validation set. For the deep learning methods, the network structures and hyperparameters were the same as described in their original paper. The learning rate was tuned in the range of [0.00005, 0.0001, 0.0005, 0.001, 0.005] on the validation sets. Then, the classification results of the test sets were generated using the optimal learning rate.

C. Classification Results

The quantitative results of all compared methods on four HSI classification datasets are shown in Tables II–V. Overall, SVM had the worst classification results among all the methods. For the CNN-based methods, SSRN achieved better results on the Salinas and Pavia University datasets, and FDSSC achieved better results on the Houston 2013 and WHU-Hi-LongKou datasets. Both of them had a substantial improvement over SVM. For the transformer-based method, the results of SpectralFormer were comparable to SVM on four datasets and significantly behind the CNN-based method. It indicated that directly training ViT on the small labeled samples was suboptimal due to the lack of inductive bias. As a solution, hybrid methods, like SSFTT, HybridFormer, DCTN, and GSC-ViT, using CNNs to enhance transformers could yield performance gains. Specifically, SSFTT and HybridFormer achieved the second-best performance on the Salinas and Pavia University datasets, respectively. DCTN achieved the second-best performance on the Houston 2013 and WHU-Hi-LongKou datasets.

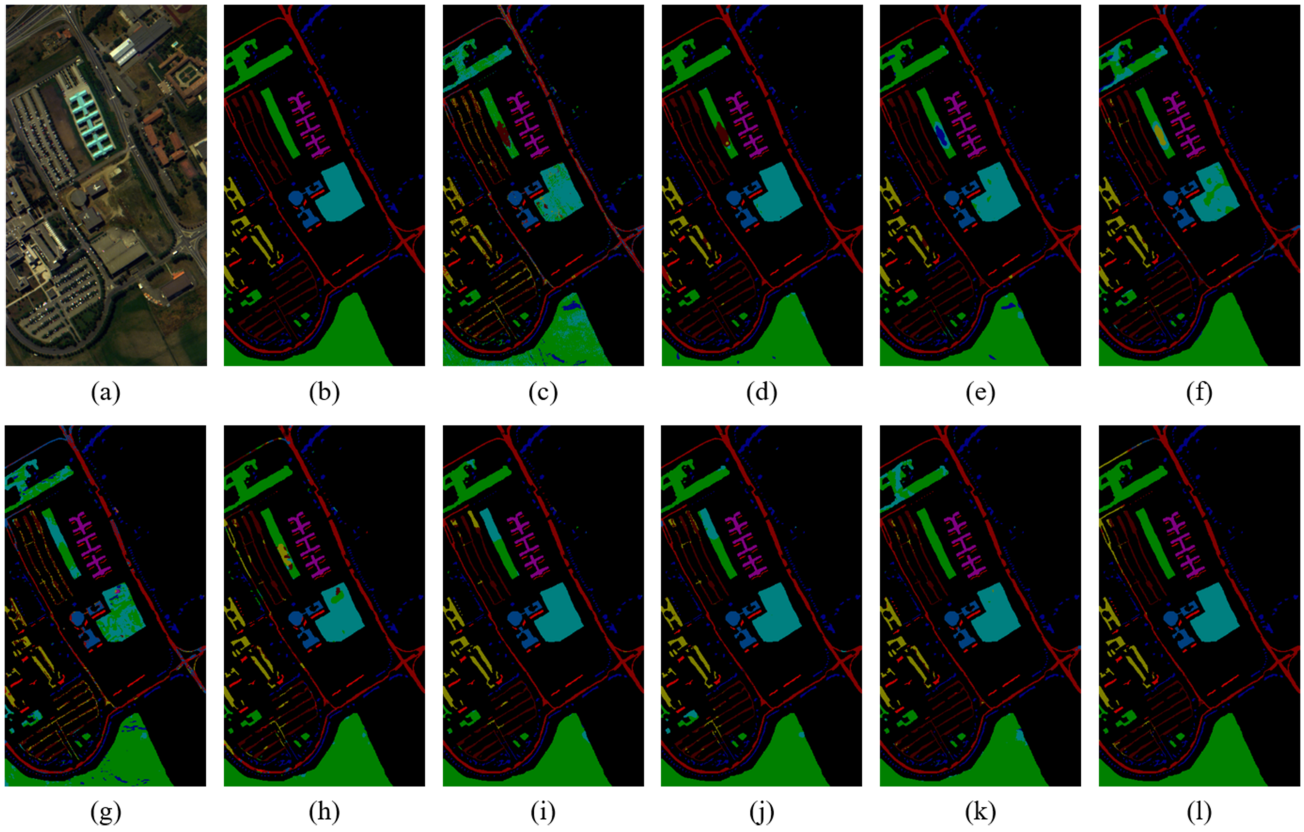

Fig. 7. Classification maps of the full Pavia University dataset. (a) Pseudocolor image. (b) Ground truth label. (c) SVM. (d) SSRN. (e)FDSSC. (f) DBDA. (g) SpectralFormer. (h) SSFTT. (i) HybridFormer. (j) DCTN. (k) GSC-ViT. (l) HSIMAE-L.

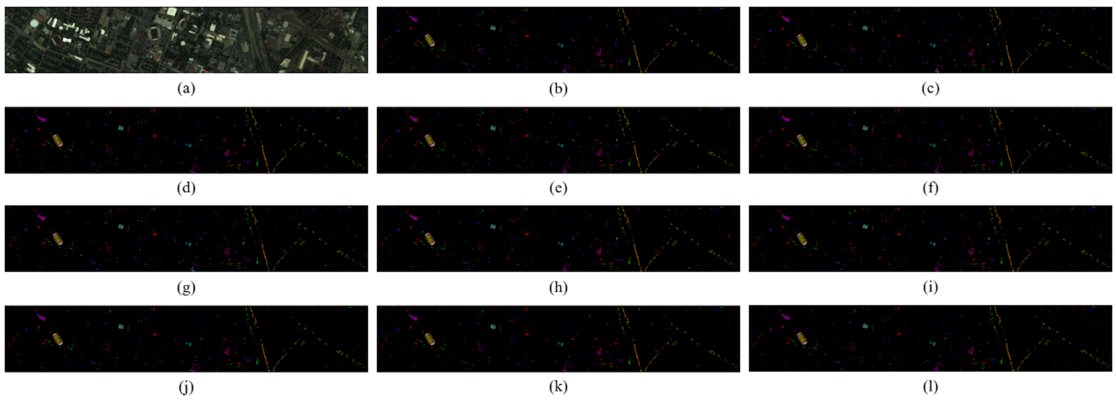

Fig. 8. Classification maps of the full Houston 2013 dataset. (a) Pseudocolor image. (b) Ground truth label. (c) SVM. (d) SSRN. (e) FDSSC. (f) DBDA. (g) SpectralFormer. (h) SSFTT. (i) HybridFormer. (j) DCTN. (k) GSC-ViT. (l) HSIMAE-L.

Instead of using elaborately modified ViT architectures, HSIMAE exploits the potential of the pure transformer by using large-scale pretraining. As can be seen in Tables II–V, where the best results of HSIMAEs are bolded, and the best results of the compared methods are underlined, HSIMAE consistently outperformed the other methods on four datasets. Specifically, HSIMAE-L achieved the highest OA of 96.62%, 97.44%, 95.65%, and 98.41%, exceeding the best compared methods by 0.25%, 1.65%, 0.50%, and 0.23% on four datasets, respectively. In particular, none of the compared methods consistently achieved second-best results, illustrating the robustness and generalization of HSIMAE.

D. Classification Maps

For further comparison, the full classification maps of different methods on the four datasets are shown in Figs. 6–9. With low classification accuracy, SVM and SpectralFormer

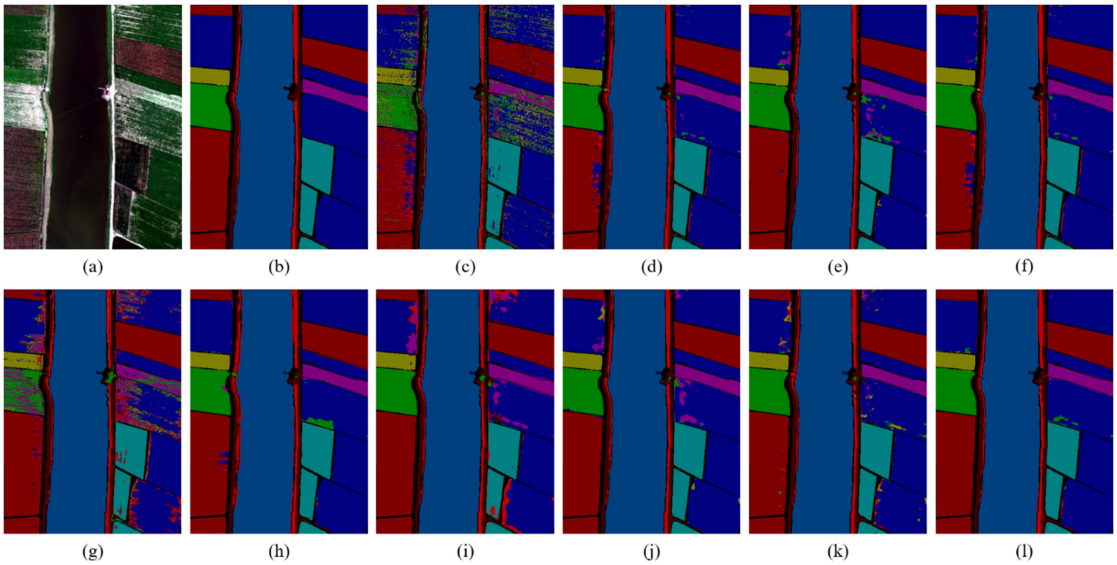

Fig. 9. Classification maps of the full WHU-Hi-LongKou dataset. (a) Pseudocolor image. (b) Ground truth label. (c) SVM. (d) SSRN. (e) FDSSC. (f) DBDA. (g) SpectralFormer. (h) SSFTT. (i) HybridFormer. (j) DCTN. (k) GSC-ViT. (l) HSIMAE-L.

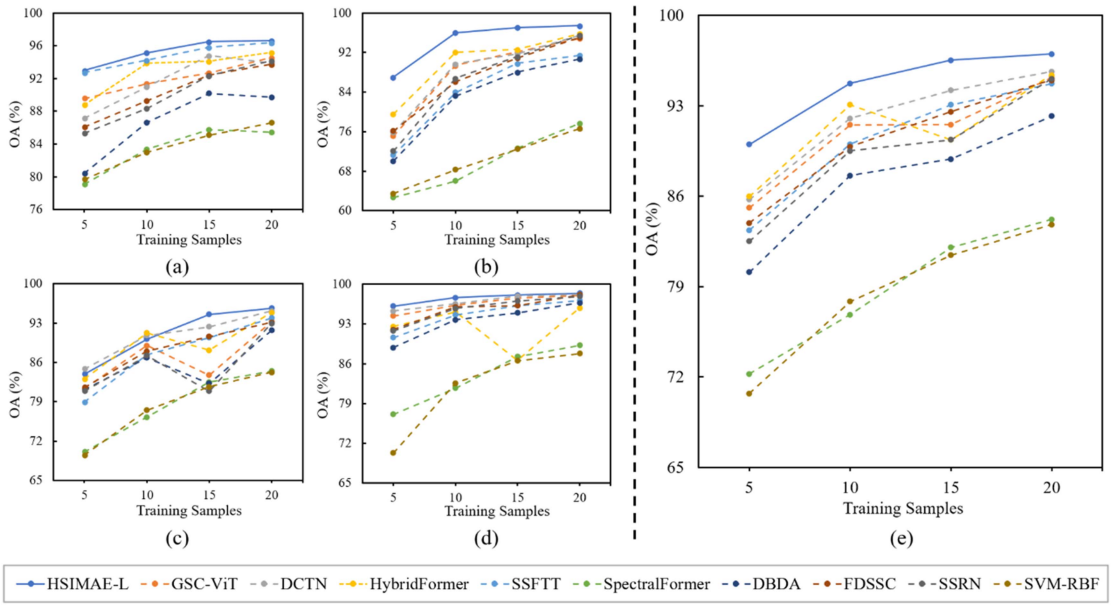

Fig. 10. Quantitative results with different numbers of training samples per class. (a) Salinas. (b) Pavia University. (c) Houston 2013. (d) WHU-Hi-LongKou. (e) Average results of the four datasets.

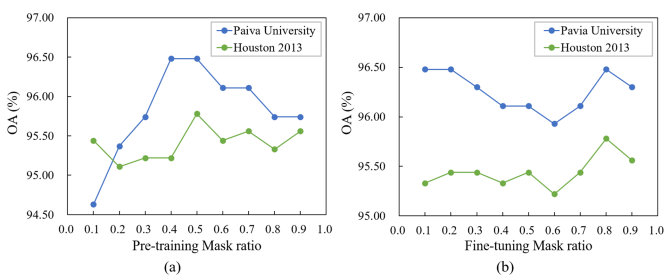

Fig. 11. Effect of the mask ratios on the classification results of the validation set. (a) Pretraining mask ratio. (b) Finetuning mask ratio.

produced many noisy regions. Despite obtaining high classification accuracy, the classification maps of CNNs lacked regional continuity and produced less visually appealing due to their limited receptive field. Owing to the long-distance modeling capabilities, the classification maps of transformers exhibited smoother results with reduced regional discontinuities compared with their CNN counterparts. Nevertheless, they still produced misjudgment regions. DCTN produced very noisy results on the Salinas dataset, which might be caused by the local features of its dual-branch convolutional transformer network dominating the classification results. Compared with other methods, the classification maps of HSIMAE-L show minimal noise

TABLE VI
RESULTS OF THE ABLATION STUDY FOR EACH COMPONENT IN THE HSIMAE FRAMEWORK

Method				Salinas	Pavia University	Houston 2013	WHU-Hi-LongKou	Average Δ
SSSE	Pre-training	Unlabeled branch						
✓	✓	✓	OA (%)	96.62	97.44	95.65	98.41	
			AA (%)	98.67	97.71	96.28	98.00	
			κ×100	96.24	96.63	95.30	97.92	
✓	✓	✖	OA (%)	95.77	96.49	95.27	97.68	-0.73
			AA (%)	98.24	96.94	96.00	97.46	-0.50
			κ×100	95.29	95.39	94.89	96.97	-0.89
✓	✖	✓	OA (%)	95.98	96.67	94.65	96.53	-1.07
			AA (%)	98.34	96.51	95.43	96.89	-0.87
			κ×100	95.53	95.60	94.21	95.49	-1.32
✖	✓	✓	OA (%)	96.39	97.19	95.56	98.13	-0.21
			AA (%)	98.59	97.14	96.26	97.87	-0.20
			κ×100	95.99	96.29	95.19	97.55	-0.27
✖	✖	✖	OA (%)	96.03	94.40	93.84	95.55	-2.08
			AA (%)	98.34	95.17	94.77	96.52	-1.47
			κ×100	95.59	92.65	93.34	94.22	-2.57

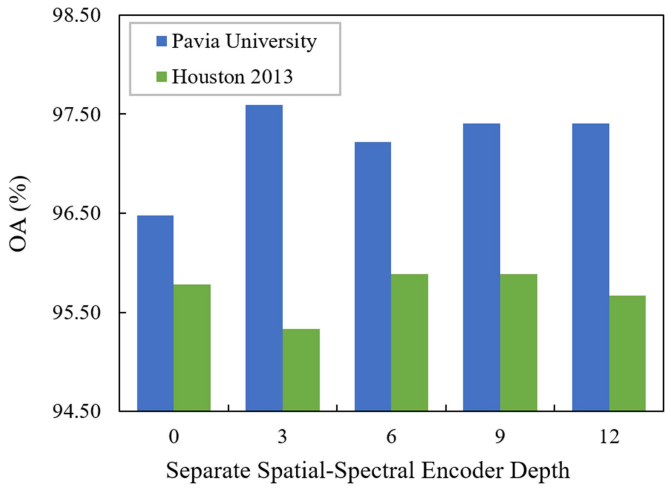

Fig. 12. Effect of the depth of separate spatial–spectral encoder on the classification results of the validation set.

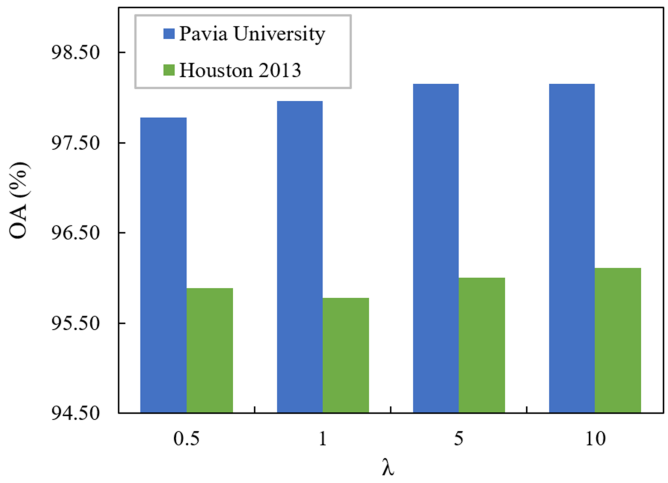

Fig. 14. Effect of the relative weight (λ) of the reconstruction loss on the classification results of the validation set.

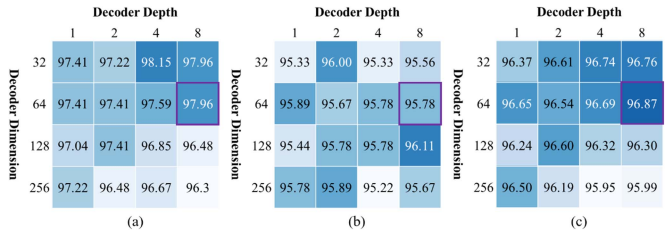

Fig. 13. Effect of the decoder capacity on the classification results of the validation set. (a) Pavia University. (b) Houston 2013. (c) Average results of the two datasets. The purple box is the default setting of HSIMAE.

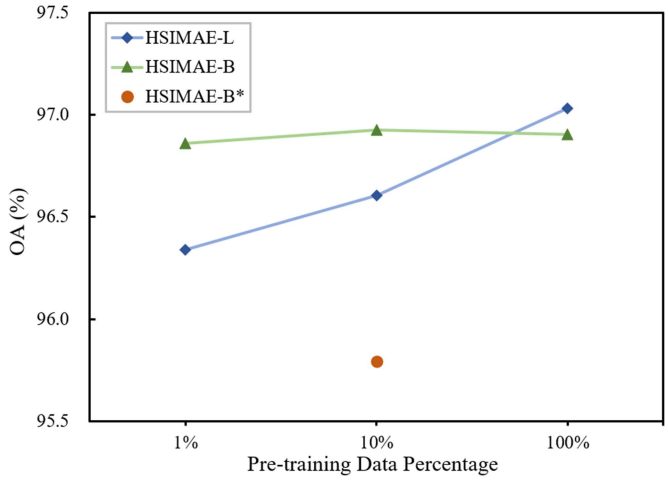

Fig. 15. Effect of the model size and dataset size on the classification results of the test set with 20 training samples. In particular, HSIMAE-B∗ used 20% of samples of the HyspecNet-11k dataset for pretraining.

regions and the best overall visualization results. It suggests that the large-scale pretraining enables the HSIMAE to learn prior information about HSI, leading to enhanced generalization capabilities.

V. Discussion

A. Impact of Training Samples

To better illustrate the robustness of the proposed method under a small number of samples, 5, 10, and 15 labeled samples per class were randomly selected as training data, and another 5, 10, and 15 random labeled samples per class were used as validation data for the four datasets. Fig. 10 shows the OAs of the compared methods with different numbers of training samples. In the case of a small number of samples, the proposed HSIMAE-L outperformed the other methods in 18 of the 20 cases, especially on the Pavia University dataset. The average OA of HSIMAE-L on the four datasets was significantly ahead of other methods, improving by 4.04%, 1.65%, 2.34% and 1.37% on 5, 10, 15, and 20 training samples, respectively. It illustrated the effectiveness and robustness of the proposed method. In summary, the classification results demonstrated that the HSIMAE framework with large- scale pre-training provides an excellent solution to build a unified HSI classification model.

B. Parameter Analysis

In this section, we analyzed the critical parameters of HSIMAE, including mask ratios for pretraining and finetuning, the depth of separate spatial–spectral encoder, the decoder design, and the relative weight of the unlabeled reconstruction loss. To improve efficiency and effectiveness of parameter selection, HSIMAE-B pretrained on 10% samples of the HSIHybrid dataset was used, and the classification results on the validation set of the Pavia University and Houston 2013 dataset with 20 samples were reported for parameter analysis.

1) *Mask Ratio:* In mask modeling, the mask ratio was an important parameter used to regulate the difficulty of the task. Here, the spatial-spectral masking strategy was used in both the pretraining and finetuning phases. As shown in Fig. 11, the mask ratio had a more significant impact on the pretraining stage than the finetuning stage. The optimal mask ratio was 0.5 and 0.8 for pretraining and finetuning, respectively. The optimal finetuning mask ratio was higher than the optimal pretraining mask ratio, which can be explained as follows: in the finetuning stage, HSIMAE was initialized by the pretrained parameters, and the model could already predict the mask part. Increasing the mask ratio forced the model to learn the feature distribution of the current dataset further. However, when using an extreme mask ratio of 90%, the reconstruction task was too difficult; thus, the performance of HSIMAE dropped.
2) *Separate Spatial–Spectral Encoders Depth:* The encoder of HSIMAE consisted of SSSE and fusion blocks. We kept the total depth of the HSIMAE encoder equal to 12. When the SSSE depth was 0, the encoder of HSIMAE was equal to the vanilla MAE encoder. When the SSSE depth increased to 12, the fusion blocks were not used and the encoder of HSIMAE became a dual-encoder structure.

As shown in Fig. 12, when the SSSE depth changes from 0 to 3, the performance of HSIMAE on the Pavia University dataset is significantly improved but decreased on the Houston 2013 dataset, which indicates that the SSSE learns different features from the original MAE encoder. As the SSSE depth increases to 9, the performance of HSIMAE suppresses the original MAE encoder on both datasets. Moreover, as the SSSE depth increases from 9 to 12, the overall accuracy of HSIMAE on the Houston dataset is dropped. This shows that both the SSSE and the Fusion blocks are essential for better performance, and the special design of the encoder of HSIMAE was useful for HSI classification.

1) *Decoder Capacity:* The effect of the decoder depth, i.e., the number of blocks and dimension of each block on the classification results, was examined. As shown in Fig. 13, the performance of HSIMAE on the Pavia University dataset was sensitive to the change in decoder capacity, and the performance difference reached 1.85%. The performance of the Houston 2013 dataset was not sensitive, and the performance difference was only 0.89%. As for the average results of the two datasets, the narrow and deep decoder was able to achieve better results compared to the shallow and wide one. This was consistent with the conclusion in MAE [20]. Therefore, we used 8-depth and 64-dimension decoder as default settings for HSIMAE.
2) *Relative Weight for Unlabeled Reconstruction Loss:* Since the unlabeled branch could be considered as a kind of regularization, which forced the model to focus on all samples in the dataset, the relative weight of the unlabeled reconstruction loss (λ) controlled the strength of the regularization. In Fig. 14, a consistent increase in OA of two datasets could be observed as the λ increase, which shows the importance of the unlabeled branch of HSIMAE in HSI classification with small labeled samples. The optimal λ was set to 10.

C. Scalability of HSIMAE

In the image and video domain, masked image modeling methods are shown to be good at scaling up model capacity [77], [78]. It makes us wonder if the scaling behavior also exists in HSIMAE. Thus, we trained HSIMAE-B and HSIMAE-H models on 1%, 10%, and 100% of samples of the HSIHybrid dataset to explore the scalability of HSIMAE and on 20% of samples of the HyspecNet-11k dataset to explore the effect of data diversity on pretraining. As shown in Fig. 5, the size of 20% of samples of the HyspecNet-11k was comparable to 10% of samples of the HSIHybrid dataset, but the data diversity of the two datasets was very different.

The average OA of each model on the four datasets is reported in Fig. 15. As the size of the pretraining dataset increases, the model performance of HSIMAE-L gradually increases, which reflects the scalability of HSIMAE models. However, the overall performance of the HSIMAE-B remains unchanged. We hypothesize that the size of the HSIMAE model should increase with the size of the pretraining dataset. The small model can achieve good performance on a small size of the pretrained dataset, but its optimal performance is lower than that of pre-trained large models. On the other hand, HSIMAE-B outperforms HSIMAE-B$*$ on the average OA by a large margin. It indicates that the diversity of the pre-training dataset is also a critical factor in better classification

performance. To summarize, HSIMAE benefits from the growth of model size and large-scale pretraining.

D. Ablation Study

To demonstrate the superiority of the HSIMAE framework, we analyzed different combinations of three core components of the framework: SSSE, larger scale pretraining, and dual-branch finetuning. When these components were discarded, HSIMAE would degenerate into simple ViT with 3-D PCA features as inputs. As shown in Table VI, each core component had its advantages, and the lack of either led to performance degradation. Among them, the larger scale pretraining had the greatest impact on the overall performance. Compared to simple ViT, HSIMAE had a significant performance gain on all four datasets, which showed the superiority of the HSIMAE framework.

VI. Conclusion

In this study, the HSIMAE framework was proposed to answer the question of "*how to train a unified HSI classification model from a large amount of unlabeled data.*" First, a large-scale HSI dataset, named HSIHybrid, was built for pretraining. It consisted of 15 HSI datasets from different hyperspectral sensors and contained a total of 4 million HSI patches. Then, to handle the different spectral resolutions and spectral ranges between HSI datasets, a group-wise PCA was used to transform the raw spectra into fixed-length features. Considering the properties of HSI data, HSIMAE used two separate spatial–spectral encoders followed by a series of fusion blocks to learn spatial correlation and spectral correlation of HSI data, and a spatial–spectral masking strategy was designed to adapt the HSIMAE encoder. Furthermore, a dual-branch finetuning framework was introduced. It used an extra unlabeled branch to adapt the model further to the distributions of the target dataset and suppress the overfitting issue. Subsequent parameter analysis experiments verified the importance of the core components of the proposed framework. As a result, even with a limited number of training samples, the HSIMAE framework outperformed existing state-of-the-art models on four classical HSI datasets from different hyperspectral sensors.

Although the proposed method performed well on HSI classification, it has not been extended to other forms of HSI analysis, such as object detection, change detection, and unmixing. Future works will apply large-scale pretrained models to the other HSI analysis tasks.

Acknowledgment

The authors are grateful for resources from the High-Performance Computing Center of Central South University.

References

[1] Q. Li et al., "Review of spectral imaging technology in biomedical engineering: Achievements and challenges," *J. Biomed. Opt.*, vol. 18, no. 10, Oct. 2013, Art. no. 100901, doi: 10.1117/1.JBO.18.10.100901.

[2] F. D. van der Meer et al., "Multi- and hyperspectral geologic remote sensing: A review," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 14, no. 1, pp. 112–128, 2012, doi: 10.1016/j.jag.2011.08.002.

[3] H. Liang, "Advances in multispectral and hyperspectral imaging for archaeology and art conservation," *Appl. Phys. A*, vol. 106, no. 2, pp. 309–323, 2012, doi: 10.1007/s00339-011-6689-1.

[4] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 010901, doi: 10.1117/1.JBO.19.1.010901.

[5] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, 2019, Art. no. 3071, doi: 10.3390/s19143071.

[6] B. Lu et al., "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens*, vol. 12, no. 16, 2020, Art. no. 2659, doi: 10.3390/rs12162659.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[9] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, Jan. 2002, doi: 10.1080/01431160110040323.

[10] 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, doi: 10.1109/TGRS.2004.842481.

[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, doi: 10.1109/TGRS.2019.2907932.

[12] T. Chao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–Spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015, doi: 10.1109/lgrs.2015.2482520.

[13] M. Zhang, M. Gong, Y. Mao, J. Li, and Y. Wu, "Unsupervised feature extraction in hyperspectral images based on Wasserstein generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2669–2688, May 2019, doi: 10.1109/tgrs.2018.2876123.

[14] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022, doi: 10.1109/mgrs.2022.3198244.

[15] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018, doi: 10.1109/tip.2017.2772836.

[16] X. Kang, B. Zhuo, and P. Duan, "Semi-supervised deep learning for hyperspectral image classification," *Remote Sens. Lett.*, vol. 10, no. 4, pp. 353–362, 2019, doi: 10.1080/2150704x.2018.1557787.

[17] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. N. Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[18] J. Zhou et al., "iBOT: Image BERT pre-training with online tokenizer," in *Proc. Int. Conf. Learn. Representation*, 2022. [Online]. Available: https://openreview.net/forum?id=ydopy-e6Dg

[19] H. Bao et al., "BEiT: BERT Pre-training of Image transformers," in *Proc. Int. Conf. Learn. Representation*, 2022. [Online]. Available: https://openreview.net/forum?id=p-BhZSz59o4

[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[21] Z. Tong et al., "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Adv. Neural Inf. Process. Syst.*, pp. 10078–10093, 2022.

[22] D. Ibanez, R. Fernandez-Beltran, F. Pla, and N. Yokoya, "Masked auto-encoding spectral–Spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5542614, doi: 10.1109/tgrs.2022.3217892.

[23] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5524312, doi: 10.1109/TGRS.2023.3315678.

[24] L. Huang, Y. Chen, and X. He, "Spectral–Spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5508718, doi: 10.1109/tgrs.2023.3264235.

[25] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–Spectral masked autoencoder for Multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5531614, doi: 10.1109/TGRS.2023.3331717.

[26] W. Liu et al., "Self-supervised feature learning based on spectral masking for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 4407715, doi: 10.1109/TGRS.2023.3310489.

[27] J. Qi et al., "Masked spatial–Spectral autoencoders are excellent hyperspectral defenders," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3345734.

[28] M. Ashraf et al., "Spatial-spectral BERT for hyperspectral image classification," *Remote Sens*, vol. 16, no. 3, 2024, Art. no. 539, doi: 10.3390/rs16030539.

[29] H. Guo and W. Liu, "S3L: Spectrum transformer for self-supervised learning in hyperspectral image classification," *Remote Sens.*, vol. 16, no. 6, 2024, Art. no. 970, doi: 10.3390/rs16060970.

[30] W. Kong, B. Liu, X. Bi, J. Pei, and Z. Chen, "Instructional mask autoencoder: A scalable learner for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1348–1362, Nov. 2024, doi: 10.1109/JSTARS.2023.3337132.

[31] S. Mohamed, M. Haghighat, T. Fernando, S. Sridharan, C. Fookes, and P. Moghadam, "FactoFormer: Factorized hyperspectral transformers with self-supervised pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Dec. 2024, Art. no. 5501614, doi: 10.1109/TGRS.2023.3343392.

[32] N. Shazeer, "GLU variants improve transformer," 2020, *ArXiv:2002.05202*.

[33] W. Hu et al., "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jul. 2015, Art. no. 258619, doi: 10.1155/2015/258619.

[34] 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, Oct. 2016, doi: 10.1109/tgrs.2016.2584107.

[35] 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, doi: 10.1109/lgrs.2017.2780890.

[36] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent Neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017, doi: 10.1109/TGRS.2016.2636241.

[37] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, Mar. 2017, Art. no. 298, doi: 10.3390/rs9030298.

[38] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 4959–4962.

[39] O. Russakovsky et al., "ImageNet large scale visual recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[40] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018, doi: 10.1109/tgrs.2018.2841823.

[41] 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, doi: 10.1109/tgrs.2018.2815613.

[42] 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, doi: 10.1109/tgrs.2017.2755542.

[43] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral–Spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017, doi: 10.1109/tgrs.2017.2698503.

[44] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2349–2361, Apr. 2018, doi: 10.1109/TGRS.2017.2778343.

[45] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6010.

[46] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representation*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[47] 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, Jan. 2020, doi: 10.1109/tgrs.2019.2934760.

[48] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2021, Art. no. 5518615, doi: 10.1109/tgrs.2021.3130716.

[49] B. Liu et al., "DSS-TRM: Deep spatial–spectral transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 103–114, 2022, doi: 10.1080/22797254.2021.2023910.

[50] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5411415, doi: 10.1109/tgrs.2022.3202036.

[51] D. Liao, C. Shi, and L. Wang, "A Spectral–Spatial fusion transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5515216, doi: 10.1109/tgrs.2023.3286950.

[52] J. Hu, B. Tu, Q. Ren, X. Liao, Z. Cao, and A. Plaza, "Hyperspectral image classification via Multi-Scale Multi-angle attention network," *IEEE Trans. Geosci. Remote Sens.*, Feb. 2024, Art. no. 5510718, doi: 10.1109/tgrs.2024.3370919.

[53] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–Spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522214, doi: 10.1109/tgrs.2022.3144158.

[54] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5528715, doi: 10.1109/tgrs.2022.3171551.

[55] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When multigranularity meets spatial–Spectral attention: A hybrid transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4401118, doi: 10.1109/tgrs.2023.3242978.

[56] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5539014, doi: 10.1109/tgrs.2022.3207933.

[57] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–Spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[58] L. Wang, Z. Zheng, N. Kumar, C. Wang, F. Guo, and P. Zhang, "Multilevel class token transformer with cross TokenMixer for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 5507913, doi: 10.1109/tgrs.2024.3361906.

[59] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, Apr. 2022, doi: 10.1109/tip.2022.3162964.

[60] F. Xu, G. Zhang, C. Song, H. Wang, and S. Mei, "Multiscale and cross-level attention learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501615, doi: 10.1109/TGRS.2023.3235819.

[61] Z. Li, Z. Xue, Q. Xu, L. Zhang, T. Zhu, and M. Zhang, "SPFormer: Self-pooling transformer for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Dec. 2024, Art. no. 5502019, doi: 10.1109/TGRS.2023.3345923.

[62] Z. Zhao, X. Xu, S. Li, and A. Plaza, "Hyperspectral image classification using groupwise separable convolutional vision transformer network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Mar. 2024, Art. no. 5511817, doi: 10.1109/tgrs.2024.3377610.

[63] X. Zhang, R. Zhang, L. Li, and W. Li, "Local–Global Cross fusion network with Gaussian-initialized learnable positional prompting for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5532216, doi: 10.1109/TGRS.2023.3335864.

[64] Y. Zhou, X. Huang, X. Yang, J. Peng, and Y. Ban, "DCTN: Dual-branch convolutional transformer network with efficient interactive self-attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 5508616, doi: 10.1109/tgrs.2024.3364143.

[65] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–Spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018, doi: 10.1109/tgrs.2017.2748160.

[66] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016, doi: 10.1109/tgrs.2015.2478379.

[67] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7758–7772, Sep. 2021, doi: 10.1109/tgrs.2020.3034133.
[68] X. Hu et al., "Contrastive learning based on transformer for hyperspectral image classification," *Appl. Sci.*, vol. 11, no. 18, 2021, Art. no. 8670, doi: 10.3390/app11188670.
[69] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, Art. no. 6008205, 2022, doi: 10.1109/lgrs.2022.3159549.
[70] 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, 2022, Art. no. 5521213, doi: 10.1109/tgrs.2021.3139099.
[71] Q. Liu et al., "Refined prototypical contrastive learning for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506214, doi: 10.1109/TGRS.2023.3257341.
[72] M. H. P. Fuchs and B. Demir, "Hyspecnet-11k: A large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods," *IEEE Int. Geosci. Remote Sens. Symp.*, pp. 1779–1782, 2023.
[73] Y. Zhong et al., "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112012, doi: 10.1016/j.rse.2020.112012.
[74] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representation*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7
[75] W. Wang et al., "A fast dense spectral–Spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1068, doi: 10.3390/rs10071068.
[76] R. Li et al., "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 582, doi: 10.3390/rs12030582.
[77] Z. Xie et al., "On data scaling in masked image modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10365–10374.
[78] L. Wang et al., "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14549–14560.

Yue Wang received the B.S. degree in applied chemistry in 2020 from the College of Chemistry and Chemical Engineering, Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in analytical chemistry.

His research interests include deep learning, pattern recognition, and hyperspectral image classification.

Ming Wen received the B.S. degree in chemistry from Henan Normal University, Xinxiang, China, in 2013, and the Ph.D. degree in analytical chemistry from Central South University, Changsha, China, in 2019.

He is currently an Engineer with the College of Chemistry and Chemical Engineering, Central South University. His research interests include deep learning, bioinformatics, and the development of innovative tools and methodologies for drug discovery.

Hailiang Zhang received the B.S. degree in applied chemistry in 2019 from the College of Chemistry and Chemical Engineering, Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in analytical chemistry.

His research interests include chemometrics, deep learning, and mass spectrometry data analysis.

Jinyu Sun received the B.S. degree in chemical engineering and technology in 2020 from the College of Chemistry and Chemical Engineering, Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in analytical chemistry.

His research interests include deep learning, and material design and screening.

Qiong Yang received the M.S. degree in organic chemistry, in 2018 from the College of Chemistry and Chemical Engineering, Central South University, Changsha, China, where she is currently working toward the Ph.D. degree in analytical chemistry.

Her research interests include deep learning, data processing, and compound identification.

Zhimin Zhang received the B.S. degree in chemical engineering and the Ph.D. degree in applied chemistry from Central South University, Changsha, China, in 2007 and 2012, respectively.

He is currently an Associate Professor with the College of Chemistry and Chemical Engineering, Central South University. His main research interests include chemometrics and cheminformatics.

Hongmei Lu received the Ph.D. degree in applied chemistry from Central South University, Changsha, China, in 2006.

She is currently a Full Professor with the College of Chemistry and Chemical Engineering, Central South University. Her main research interests include hyperspectral image processing, chemometrics, and cheminformatics.