S^2HM^2: A Spectral–Spatial Hierarchical Masked Modeling Framework for Self-Supervised Feature Learning and Classification of Large-Scale Hyperspectral Images

Lilin Tu, Jiayi Li, *Senior Member, IEEE*, Xin Huang, *Senior Member, IEEE*, Jianya Gong, Xing Xie, and Leiguang Wang

***Abstract*— Most of the existing deep learning-based hyperspectral image (HSI) classification algorithms are based on supervised learning, where a large number of annotated labels with high acquisition cost are required. Self-supervised learning (SSL) methods can learn abundant representations using a large amount of unlabeled data, thereby reducing the reliability of labels. In particular, SSL based on masked image modeling (MIM) can extract fine-grained features, which is well suited for HSI classification as a pixel-level interpretation task. However, MIM has scarcely been investigated in the HSI classification. Current algorithms lack a comprehensive consideration of the multiscale spectral–spatial characteristics of HSI when constructing the pretraining task, and there exists high computational cost and redundancy when applied to large-scale HSIs. Therefore, this article develops an SSL framework based on spectral–spatial hierarchical masked modeling (S^2HM^2) for large-scale HSI classification. Considering the spectral–spatial characteristics of HSI, 3-D masking strategy and spectral–spatial consistency loss are proposed to construct the MIM task. To fully exploit features at each scale, hierarchical 3-D feature pyramid network (3D-FPN) is designed as decoder for both pretext and downstream tasks in a "pixel-to-pixel" manner. In addition, multiscale masked feature modeling (MS-MFM) task is proposed to further facilitate the multiscale feature learning. The SSL pretraining is guided by both MIM and MS-MFM. The experimental results on two large-scale hyperspectral datasets, i.e., WHU-OHS and WHU-H^2SR, demonstrate the superiority of the proposed method. Furthermore, transfer learning experiments are conducted on a variety of hyperspectral datasets, where classification accuracies are boosted in most of the scenarios. The source code will be made available at https://github.com/tulilin/S2HM2.**

Manuscript received 6 February 2024; revised 12 April 2024; accepted 20 April 2024. Date of publication 24 April 2024; date of current version 8 May 2024. This work was supported in part by the Major Scientific and Technological Projects of Yunnan Province under Grant 202202AD080010; and in part by the National Natural Science Foundation of China under Grant 42071311, Grant 42271328, and Grant 42090011. *(Corresponding author: Jiayi Li.)*

Lilin Tu, Jiayi Li, Xin Huang, and Jianya Gong are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: tulilin0312@163.com; zjjercia@whu.edu.cn; xhuang@whu.edu.cn; gongjy@whu.edu.cn).

Xing Xie is with the School of Environmental and Mapping Engineering, Suzhou University, Suzhou 234000, China.

Leiguang Wang is with the Institute of Big Data and Artificial Intelligence and the Key Laboratory of State Forestry and Grassland Administration on Forestry and Ecological Big Data, Southwest Forestry University, Kunming 650024, China.

Digital Object Identifier 10.1109/TGRS.2024.3392962

***Index Terms*— Hyperspectral image (HSI) classification, masked image modeling (MIM), multiscale features, self-supervised learning (SSL), spectral–spatial information.**

I. Introduction

HYPERSPECTRAL remote sensing, which can obtain rich spectral information of land surface, is one of the most important Earth observation technologies [1], [2]. Hyperspectral image (HSI) classification, i.e., assigning a semantic label to each pixel, is one of the key techniques for HSI analysis [3]. Large-scale HSI classification has a wide range of applications, e.g., mineral exploration [4], precision agriculture [5], environment monitoring [6], and urban planning [7].

With the advantage of extracting high-level semantic representation of data, deep learning has achieved unparalleled progress in HSI classification [8], [9]. Most of the existing deep learning-based methods are implemented in a supervised manner, which requires a large amount of labeled data to train the deep models [10], [11], [12], [13], [14]. However, the semantic annotation of HSI is time-consuming and expensive [15], which severely restricts the performance of HSI classification. To address this issue, self-supervised learning (SSL) is a new learning paradigm developed in recent years. Through generating supervised signals from data itself to guide the network training [16], SSL can learn abundant representations by leveraging the large amount of unlabeled data [17]. In this scenario, with only a few labels for fine-tuning the network after SSL pretraining, a satisfactory performance can be achieved [18], [19]. SSL can effectively resolve the problem of label scarcity and has great potential for HSI classification.

Contrastive learning [20], [21], [22], [23] methods are widely adopted for SSL and have been applied to HSI classification [24]. The goal of contrastive learning is to pull together the representations of different views of the same image (i.e., positive samples) and push apart those of different images (i.e., negative samples) [25]. It should be noticed that effective semantic representations are able to separate objects of different classes and gather those of the same

class. However, while contrastive learning can capture the difference between multiple land cover classes, it fails to model the intraclass diversity [26]. In the meantime, HSI classification is a pixel-level interpretation task, where assigning effective semantic information to each pixel is of great importance. In recent years, SSL based on masked image modeling (MIM) [27], [28] has drawn unprecedented attention in computer vision [29]. MIM aims at reconstructing the masked parts of an image using the visible parts via an encoder–decoder network. In this process, encoder is expected to learn high-level semantic representations of images [30]. Compared with discriminative contrastive learning, MIM is based on pixel-level reconstruction task. Therefore, MIM can perceive the spatial details in the images and learn fine-grained semantic representations, which are more suitable for HSI classification. Although MIM-based SSL has been continuously developing in remote sensing [31], [32], [33], [34], the research in the field of HSI classification is scarce. Current research mainly exists the following issues.

1) In view of decoder, current research does not fully utilize the low-level features extracted by the shallow encoder layers [35] for image reconstruction. Thus, it is therefore unsuitable for the task of HSI classification, which concentrates on multiscale representations [19].
2) Most of the existing MIM-based SSL algorithms generate masks for HSI in either spectral [36] or spatial [37] dimension without a comprehensive consideration of the spectral–spatial features. In addition, these networks are optimized by minimizing the pixel-wise difference between reconstructed and original images, lacking the considering from continuous spectral curves and spatial structure of HSI.
3) Most of the existing algorithms generate overlapping patches using a window of small size as input. When these methods are applied to large-scale HSI, it is difficult for networks to learn long-range relationships in the image given the limited spatial window [38]. Moreover, there exists unacceptable high computational cost and redundancy for the patch-wise processing.

Therefore, in this article, we propose a pixel-wise SSL framework based on spectral–spatial hierarchical masked modeling (S^2HM^2) for large-scale HSI classification. The main contributions are as follows.

1) To make full use of the encoding features at each scale, a 3-D feature pyramid network (3D-FPN) is designed as decoder. In 3D-FPN, the decoding features are gradually upsampled and fused with the encoding features at current scale, to recover the spatial information and further implement "pixel-to-pixel" image reconstruction (for SSL pretraining) or HSI classification (for the downstream task).
2) To further facilitate the multiscale feature extraction of HSI, the multiscale masked feature modeling (MS-MFM) task is proposed, where encoding features at each scale are reconstructed using the corresponding decoding features. By introducing the target encoder with momentum update, the quality of the generated reconstruction targets is improved, thus boosting the performance of multiscale feature learning.
3) Considering the spectral–spatial characteristics of HSI, a spectral–spatial 3-D masking strategy and a spectral–spatial consistency loss are proposed to construct the MIM task. Three-dimensional masks are generated for HSI in both spectral and spatial dimensions, and a uniform loss function that measures the reconstruction of spectral curves and preserves the spatial details is developed. The SSL pretraining is guided by the combination of the spectral–spatial MIM and MS-MFM.

II. Related Works

A. Supervised HSI Classification Based on Deep Learning

Deep learning approaches have made great advances in supervised HSI classification over the past years [8], [9]. Among the various deep networks, convolutional neural network (CNN) can effectively extract the contextual information in the images via the sparse connection and weight-sharing mechanisms and is one of the most popular network structures for HSI classification [8]. In [10], CNN with 3-D convolution kernels was proposed for the spectral–spatial feature extraction of HSI. A pyramidal residual network architecture was proposed in [39] to increase the diversity of high-level spectral–spatial features. The unified multiscale learning (UML) framework [40] adopted convolutional layers with different dilation rates for multiscale feature learning of HSI, and the features were further enhanced by channel shuffling operation and spatial–spectral attention. In recent years, Transformer [41], [42] has been successfully applied in the field of computer vision [43]. Compared with CNN which extracts features in local receptive field, Transformer can capture the global dependency in the images through multihead self-attention (MSA). Several Transformer-based networks have been developed for HSI classification. In [13] and [44], the spectral bands of HSI were groupwise embedded as the inputs of the MSA. The convolution operations were incorporated into Transformer in [14] to capture the subtle spectral–spatial discrepancies. A hybrid network structure combining CNN and Transformer was proposed in [45], where the patch attention module was integrated with convolution and Transformer blocks to extract global–local features. For HSI classification, both local and global features are important [14], [46]. Therefore, in this article, the network structure was designed taking both CNN and Transformer into consideration.

B. Masked Image Modeling

Motivated by the success of masked language modeling (MLM) [47], [48] in the field of nature language processing (NLP), in recent years, SSL based on MIM has developed rapidly and made great breakthrough in computer vision [29]. These methods mask some parts of the images and learn visual representations by recovering the missing information using the visible parts.

The information density of images is sparse and there exists large amount of redundancy. In this case, the missing

pixels within every image can easily be recovered using the neighborhood information. Therefore, to achieve a high-level understanding on the images rather than learning "short-cut" solutions, a large portion of the images can be masked to build the reconstruction task. Based on the above considerations, masked autoencoders (MAEs) [27] was proposed as the first milestone for MIM. In MAE, images were first split into nonoverlapping patches, and a high ratio of patches were randomly chosen to be masked. Only visible patches were sent into an encoder to extract the latent features of the whole image. Locations of the masks and the latent features were fed into a decoder to estimate each masked patch. Furtherly, a simple MIM framework named SimMIM [28] was proposed. Different from MAE, after the random masking, the masked patches were replaced by learnable tokens and processed by the encoder–decoder network together with the visible patches. In this way, the 2-D structure of the images was preserved throughout the learning process, which allows SimMIM to be flexible in the selection of encoder network structures, especially suitable for hierarchical networks (e.g., Swin Transformer [49]). Hierarchical networks can extract multiscale features, which is important for HSI classification. Therefore, in this article, we follow the practice of SimMIM to design the SSL framework for feature learning and classification of large-scale HSI.

C. HSI Classification Based on SSL

For HSI classification based on SSL, deep networks are first pretrained to learn representations from images and then transferred to the downstream HSI classification task.

The SSL methods for HSI classification are mainly based on contrastive learning. For most of current studies, different views of HSI are generated in spectral or spatial domain to construct the positive and negative samples, and networks are trained with the goal of minimizing the distance between positive samples and maximizing the distance between negative ones. For example, deep multiview learning method [50] divided the spectral bands of HSI into two groups as different views. In [51], different views of HSI in spectral and spatial domains were generated through Gaussian noise and spatial transformations, respectively, and contrastive learning was implemented after merging spectral–spatial features. A 3-D Swin Transformer (3DSwinT) network structure and a hierarchical contrastive learning method [19] were proposed for multiscale spectral–spatial and global–local feature extraction. Supervised contrastive learning [52], which regards samples from the same (different) classes as positive (negative) samples, were introduced to HSI classification with limited labeled samples in [53]. In this framework, pseudo-labels were generated based on spectral–spatial mixing distance to improve the contrastive learning. In [54], refined prototypical contrastive learning was proposed for few-shot HSI classification, where supervised contrastive learning was served as one of the constraints on the prototypes. In addition, contrastive learning based on BYOL [22], which using only positive samples, has also been applied to HSI classification [55]. For these methods, the targets for contrastive learning are generated by a momentum encoder, which stabilizes the network training and improves the quality of learned representations. Inspired by BYOL, in this article, target encoder with momentum update is introduced to generate the reconstruction targets of MS-MFM task.

SSL algorithms based on MIM are scarcely investigated in the field of HSI classification. For the reconstruction of spectral information, based on MAE, MAEST [36] conducted groupwise spectral embedding for HSI, where a portion of band groups were randomly selected to be masked. The pretext task of MAEST refers as to reconstruct the masked spectral bands. For the reconstruction of spatial information, spectral–spatial masked Transformer (SS-MTr) [37] masked HSI in the spatial dimension to build the reconstruction task. In addition, SS-MAE [56] is proposed for the joint classification of HSI and light detection and ranging (LiDAR) [or synthetic aperture radar (SAR)] data, which includes a spatial branch reconstructing the masked spatial patches and a spectral branch reconstructing the masked spectral channels. In this article, an SSL framework based on spectral–spatial 3-D Mask and hierarchical 3-D decoding is proposed for large-scale HSI classification.

III. Methodology

A. Overview

An SSL framework based on S^2HM^2 for large-scale HSI classification is proposed in this study. The overall structure of S^2HM^2 is illustrated in Fig. 1, which consists of the following steps.

Step 1 (Spectral–Spatial 3-D Masking and Feature Extraction of HSI): Considering the abundant spectral–spatial information, 3-D Mask strategy is performed on the HSI, where a portion of patches in both spectral and spatial dimension are selected and masked. The masked HSI is then fed to the 3DSwinT [19] encoder to extract spectral–spatial features.

Step 2 (Hierarchical 3-D Decoding): To make full use of the multiscale features, 3D-FPN is designed as the decoder for both SSL pretraining and the downstream HSI classification task. In 3D-FPN, the features outputted by the last encoder layer are gradually upsampled and fused with encoding features at current scale to recover the spatial details.

Step 3 (Image Reconstruction With Spectral–Spatial Consistency and Multiscale Feature Reconstruction): The masked parts of the input HSI are reconstructed from the output features of 3D-FPN. To model the characteristics of spectral curves and spatial structure, reconstruction loss based on spectral–spatial consistency is developed for training the MIM task. In addition, MS-MFM task is proposed that reconstructing the encoding features at each scale from the corresponding decoding features, where the reconstruction targets are generated by introducing a target encoder with momentum update. The full pretext task is the combination of the spectral–spatial MIM and MS-MFM.

B. Spectral–Spatial 3-D Masking and Feature Extraction

1) 3-D Mask: Existing MIM-based methods generate masks of HSI in only spectral or spatial dimension (denoted as

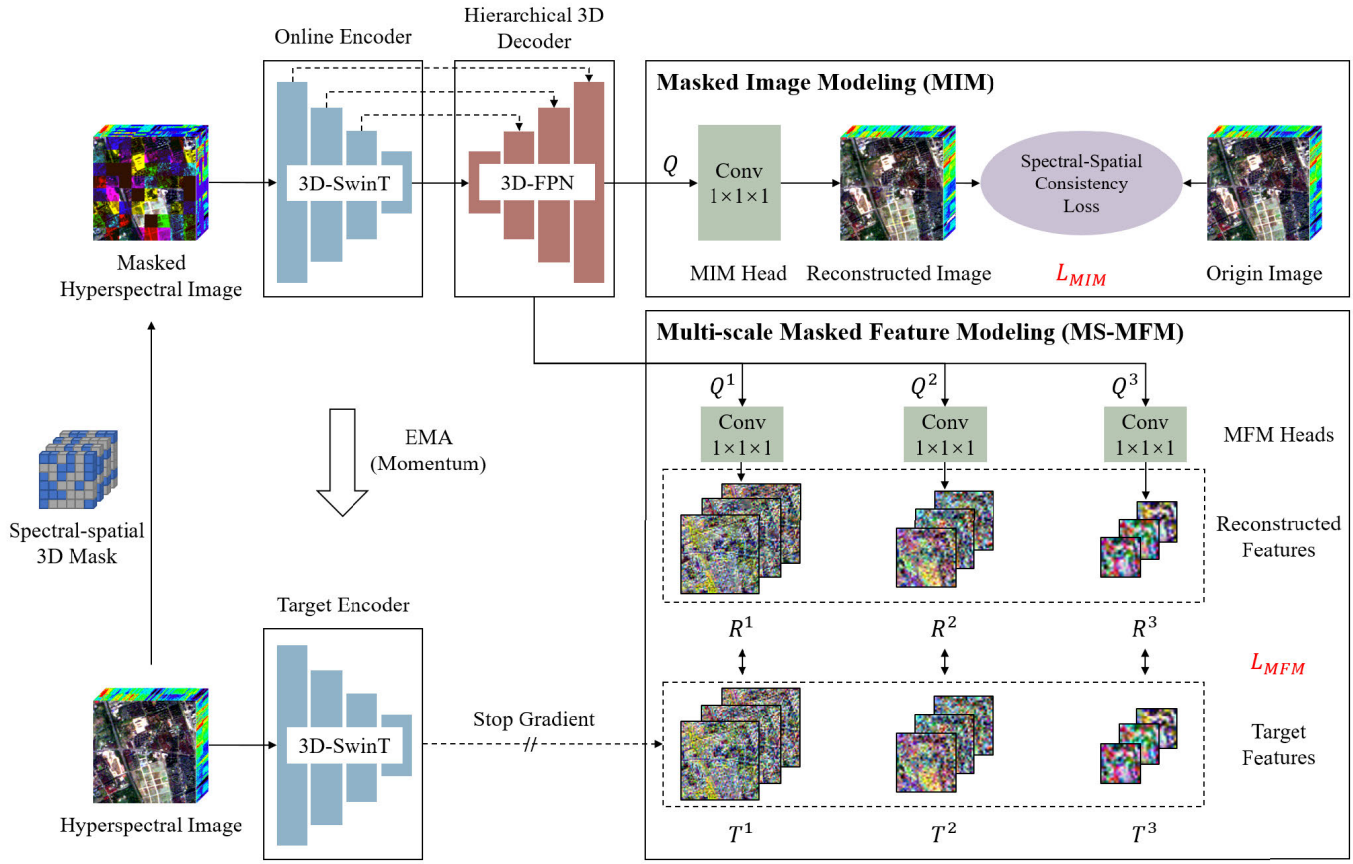


Fig. 1. Overall structure of the proposed S^2HM^2. Given an HSI, 3-D Mask strategy is performed to mask the patches selected in both spectral and spatial dimension. The masked image is then fed into 3DSwinT encoder to extract multiscale spectral–spatial features. The encoding features are hierarchically decoded using the 3D-FPN and the masked patches are reconstructed. The network is jointly trained by the MIM task based on spectral–spatial consistency loss, and the MS-MFM task with the reconstruction targets generated by a momentum target encoder.

1-D Mask and 2-D Mask, respectively). In this article, 3-D Mask strategy is proposed. Through masking HSI in both spectral and spatial dimensions, the network can achieve a more comprehensive understanding of the spectral–spatial characteristics.

For an HSI X with the size of $B \times H \times W$ (H and W denote the height and width of the image, respectively, and B is the number of spectral bands), it is first split into several 3-D cubes X^i with the size of $b \times h \times w$

$$X = \{X^i\}_{i=1}^{N}, \quad \text{where} \quad N = \frac{B}{b} \times \frac{H}{h} \times \frac{W}{w}. \tag{1}$$

Given α within the range of (0, 1), the cubes with the ratio of α are randomly selected and masked

$$\text{index} = \text{RandomSelect}(\alpha N, N) \tag{2}$$

$$M = \{M^i\}_{i=1}^{N} \tag{3}$$

$$M^i = \begin{cases} 1, & i \text{ in index} \\ 0, & \text{other} \end{cases} \tag{4}$$

where M is the obtained mask for the whole HSI, with the size of $B \times H \times W$. Equation (2) randomly picks up the cube indexes among the integers from 1 to N with the ratio of α. $M^i = 1$ represents that the cube is masked, while $M^i = 0$ indicates that the pixels with number of $b \times h \times w$ in the cube are visible during the training process.

Fig. 2 shows the comparison between 1-D Mask, 2-D Mask, and the proposed 3-D Mask. The 1-D Mask and 2-D Mask consider either spectral or spatial dimension of HSI, leading to information loss in the other dimension and hindering effective feature extraction. For example, with respect to 1-D Mask [Fig. 2(a)], a portion of spectral bands are selected and all the spatial regions of the selected bands are masked, which makes it difficult to reconstruct the spatial context in the masked bands. Similarly, the 2-D Mask strategy [Fig. 2(b)] masked all spectral bands for the selected spatial regions, increasing the difficulty of spectral curve reconstruction. In contrast, the proposed 3-D Mask [Fig. 2(c)] takes the redundancy in both spectral and spatial dimensions into consideration, which can improve the capability for the network to learn spectral–spatial representations.

2) Feature Extraction of HSI: In this article, 3DSwinT [19] is selected as the network structure of the encoder considering its capacity to capture the multiscale spectral–spatial information of HSI.

In 3DSwinT, the HSI is first split into 3-D patches with the size of $d \times 4 \times 4$ and feature with the size of $C \times D \times (H/4) \times (W/4)$ is obtained (i.e., C feature maps with the size of $D \times (H/4) \times (W/4)$, where $D = B/d$). Then, according to the 3-D Mask generated in Section III-B1, the masked patches are replaced by learnable

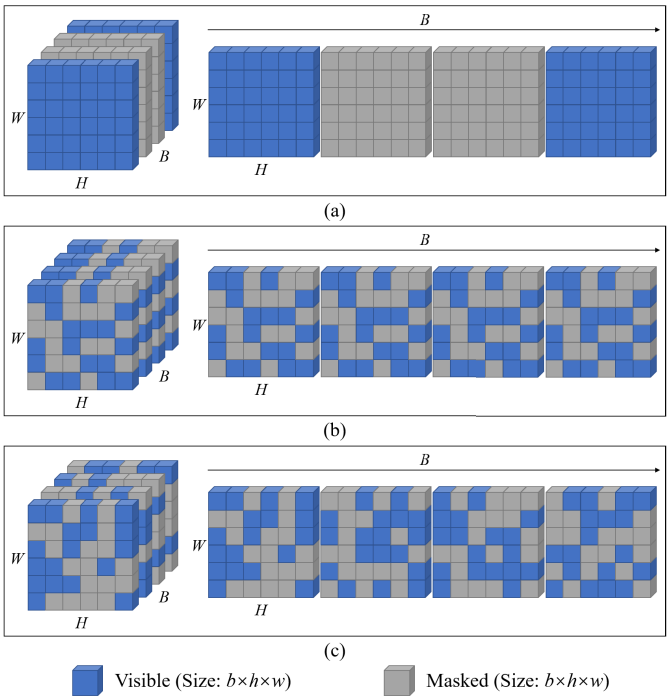


Fig. 2. Comparison between (a) 1-D Mask, (b) 2-D Mask, and (c) proposed 3-D Mask. In every subfigure, there are $(B/b) \times (H/h) \times (W/w) = 144$ 3-D cubes, and the mask ratio α is 0.50.

masked tokens with random initialization as the subsequent input.

The main components of 3DSwinT include four blocks. Each block consists of several pairs of 3-D window-based multihead self-attention (3-D W-MSA) and 3-D shifted window-based MSA (3-D SW-MSA). The 3-D W-MSA extracts the spectral and spatial sequence information in each local 3-D window, and the 3-D SW-MSA exchanges information between windows. In this way, the spectral–spatial features of HSI can be captured locally and globally. In addition, neighboring patches are merged between each of the two adjacent blocks to construct multiscale feature maps [19]. Finally, the network extracts the features of HSI with four different scales denoted as P^1–P^4, and the feature sizes are $C \times D \times (H/4) \times (W/4)$, $2C \times D \times (H/8) \times (W/8)$, $4C \times D \times (H/16) \times (W/16)$, and $8C \times D \times (H/32) \times (W/32)$.

C. Hierarchical Decoding Based on 3D-FPN

The masked regions of HSI are reconstructed based on the features extracted by the encoder. In this article, a feature pyramid network with 3-D convolution is designed for the hierarchical decoding of both SSL pretraining and the downstream HSI classification. In the process of decoding, the low spatial resolution with high-level semantic features is gradually upsampled and fused with the high spatial resolution low-level encoding features at current scale [57]. In this way, the features at each scale are fully utilized, and the image reconstruction or HSI classification can be effectively implemented in a "pixel-to-pixel" manner. The network structure of 3D-FPN is depicted in Fig. 3.

The output feature P^4 of the last encoder layer is fed into a 3-D feature aggregation module (3D-FAM), as shown in Fig. 3(b). The adaptive 3-D max pooling with four different sizes is applied to P^4, and feature maps with the size of $1 \times 1 \times 1$, $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $6 \times 6 \times 6$ are obtained. These feature maps are then upsampled to the size of P^4 and concatenated together after being processed by a $1 \times 1 \times 1$ convolution. The output feature Q^4 of the 3D-FAM, with the size of $C' \times D \times (H/32) \times (W/32)$ (C' is the number of output feature maps in 3D-FPN), is obtained using another $1 \times 1 \times 1$ convolution. In a word, 3D-FAM further enlarges the receptive field of the network through aggregating the contextual information in different regions [58].

The feature Q^4 is then passed through three consecutive 3D-FPN blocks, and each one is shown in Fig. 3(c). The input feature Q^n at the nth scale is upsampled with the ratio of 2, and the encoding feature P^{n-1} at corresponding scale is fed into a 3-D convolution block (composed of $1 \times 1 \times 1$ convolution, 3-D batch normalization, and rectified linear unit (ReLU) activation function) to unify the feature size. These two features are added together and forwarded to another 3-D convolution block to get the output of 3D-FPN block Q^{n-1}, which is also the input feature at the $(n-1)$th scale. After going through three 3D-FPN blocks, decoding features at three different scales Q^3, Q^2, and Q^1 are obtained, which are subsequently used for the MS-MFM task. The feature sizes are $C' \times D \times (H/16) \times (W/16)$, $C' \times D \times (H/8) \times (W/8)$, and $C' \times D \times (H/4) \times (W/4)$, corresponding to the encoding features P^3, P^2, and P^1.

Finally, features Q^4, Q^3, Q^2, and Q^1 are upsampled to unify the feature size and concatenated together. A 3-D convolution block is adopted for fusing these multiscale features to obtain the output feature Q, with the size of $C' \times D \times (H/4) \times (W/4)$, which is used to reconstruct the original image (i.e., the proposed spectral–spatial MIM task).

D. Image Reconstruction With Spectral–Spatial Consistency and Multiscale Feature Reconstruction

1) MIM Based on Spectral–Spatial Consistency: A $1 \times 1 \times 1$ convolution is employed as reconstruction head to obtain the reconstructed image from the output feature Q of 3D-FPN. The size of original image is $B \times H \times W$ and the size of feature Q is $C' \times D \times (H/4) \times (W/4)$. Therefore, the number of output feature maps for the reconstruction head should be $16d = d \times 4 \times 4$ (as mentioned before, $D = B/d$), which represents the values of d bands for 4×4 pixels. In this way, the output of the reconstruction head [with the dimension of $16d \times D \times (H/4) \times (W/4)$] can be transformed to the original image. The detailed steps of the transformation are as follows.

Step 1: Reshape the output feature to the size of $d \times 4 \times 4 \times D \times (H/4) \times (W/4)$ by unflattening the first dimension.

Step 2: Permute the dimensions of the feature to the size of $D \times d \times (H/4) \times 4 \times (W/4) \times 4$.

Step 3: Reshape the feature to the size of $B \times H \times W$ by merging each of the two dimensions.

Existing algorithms mostly adopt the $L1$ loss (i.e., the pixel-wise difference between the reconstructed and original

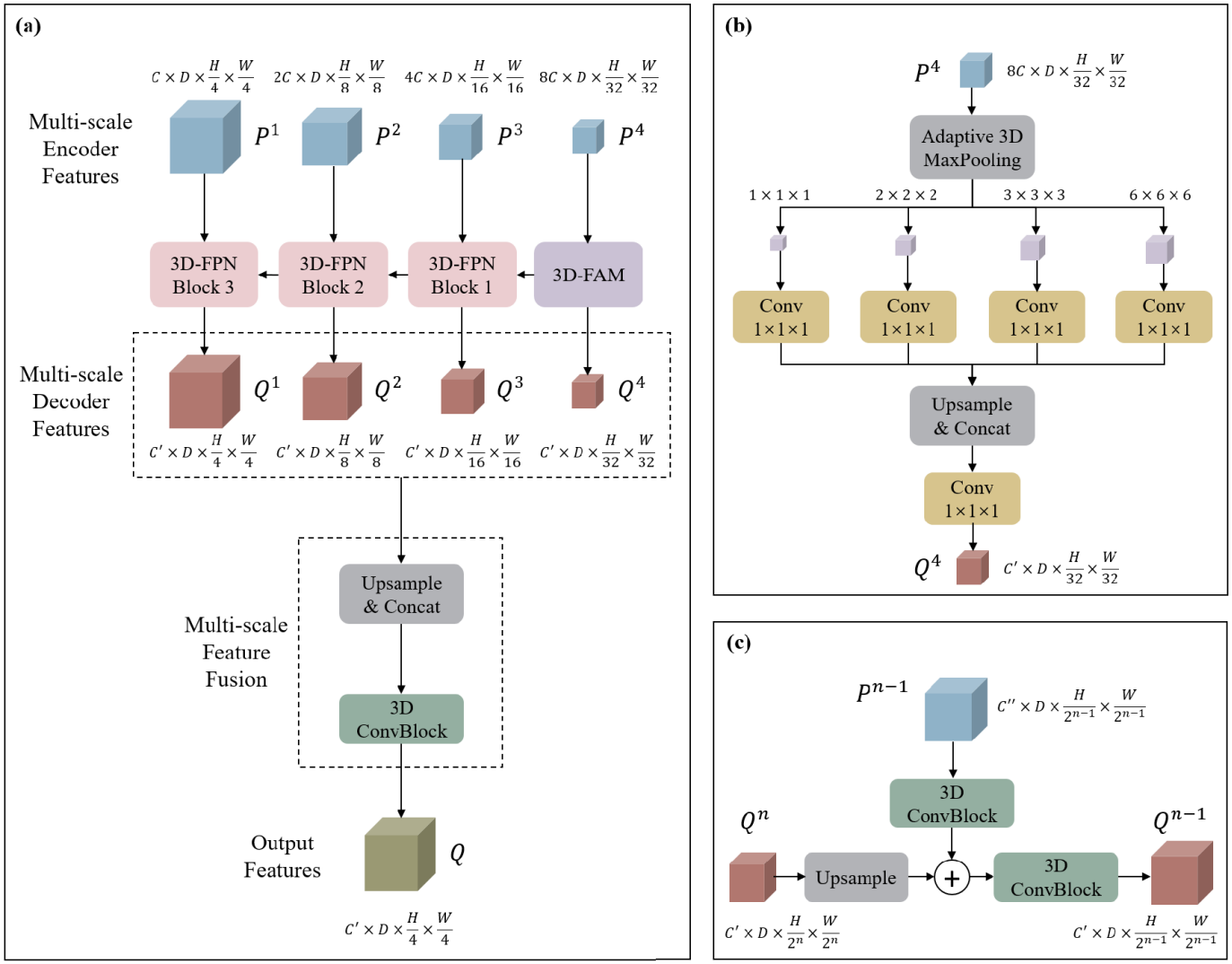


Fig. 3. Network structure of 3D-FPN. (a) Overall structure. (b) 3D-FAM. (c) 3-D feature pyramid network (3D-FPN) block, where C'' is the number of feature maps in the encoding features at current scale.

image) for the training of MIM task

$$L_1 = \frac{\sum M \odot |X_{\text{rec}} - X|}{\sum M} \tag{5}$$

where X_{rec} and X represent the reconstructed and original images, respectively; M is the mask; and $\odot$ represents the element-wise product.

To better model the characteristics of spectral curves and spatial structure in HSI, in this article, spectral–spatial consistency loss is designed for the optimization of the spectral–spatial MIM task based on spectral angle mapper (SAM) and structure similarity index measure (SSIM). The loss function based on SAM measures the cosine similarity of the reconstructed and original spectral curves [59]

$$L_{\text{SAM}} = \frac{1}{HW} \sum_{i=1}^{HW} \frac{x_i^T x_i'}{||x_i||\,||x_i'||} \tag{6}$$

where x_i' and x_i represent each pixel of reconstructed and original images, respectively.

The loss function based on SSIM evaluates the expression of spatial details in the image

$$\text{SSIM} = \frac{(2\mu_p\mu_q + C_1)(2\sigma_{pq} + C_2)}{(\mu_p^2 + \mu_q^2 + C_1)(\sigma_p^2 + \sigma_q^2 + C_2)} \tag{7}$$

$$L_{\text{SSIM}} = \frac{\sum M \odot (1 - \text{SSIM})}{\sum M} \tag{8}$$

where (7) is the calculation of SSIM [60]. p and q represent the reconstructed and original images, respectively. μ and σ are the mean and standard deviation of the neighborhoods of each pixel, respectively. C_1 and C_2 are set to 0.01^2 and 0.03^2, respectively, to avoid 0 in denominator [60], [61].

The proposed spectral–spatial consistency loss for MIM can be expressed as

$$L_{\text{MIM}} = L_1 + \lambda_1 L_{\text{SAM}} + \lambda_2 L_{\text{SSIM}} \tag{9}$$

where λ_1 and λ_2 are the weight hyperparameters for SAM and SSIM losses, respectively.

TABLE I
HYPERSPECTRAL DATASETS USED FOR EXPERIMENTS

Dataset	Spatial resolution (m)	Spatial Coverage (km^2)	Data size (pixels)	Bands	Wavelength (nm)	Classes
WHU-OHS [3]	**10.0**	**204341.25**	**7795×512×512**	**32**	**466-940**	**24**
WHU-H^{2}SR [64]	**1.0**	**227.79**	**2531×300×300**	**249**	**391-984**	**8**
Indian Pines	20.0	8.41	145×145	200	400-2500	16
Pavia University	1.3	0.35	610×340	103	430-850	9
Pavia Center	1.3	1.32	1096×715	102	430-850	9
Salinas	3.7	1.52	512×217	204	400-2500	16
KSC	18.0	101.86	512×614	176	400-2500	13
Botswana	30.0	340.07	1476×256	145	400-2500	14
DFC2013	2.5	4.16	349×1905	144	380-1050	15
DFC2018	1.0	5.75	4786×1202	50	380-1050	20
Washington DC	1.5	0.88	1280×307	191	400-2400	7
AeroRIT [65]	0.4	1.25	1973×3975	51	400-900	5
WHU-Hi-HanChuan [66]	0.109	0.004	1217×303	274	400-1000	16
WHU-Hi-HongHu [66]	0.043	0.001	940×475	270	400-1000	22
WHU-Hi-LongKou [66]	0.463	0.047	550×400	270	400-1000	9
Matiwan [67]	0.5	1.48	1580×3750	256	400-1000	20

2) Multiscale Masked Feature Modeling: To further extract the multiscale features of HSI, multiscale mask feature modeling (MS-MFM) task is proposed in this article, where the decoding features at each scale are utilized to reconstruct the corresponding encoding features. We denote the encoder in Section III-B as online encoder and introduce another symmetric encoder named target encoder to generate the reconstruction targets. The parameters of the target encoder are updated using exponentially moving average (EMA) according to the parameters of the online encoder (i.e., momentum update with stop gradient)

$$\theta_i^t = m\theta_{i-1}^t + (1-m)\theta_i^o \tag{10}$$

where θ^t and θ^o represent the parameters of target encoder and online encoder, respectively; i is the current training steps; and m is a relative high momentum value (set to 0.996 [22] in this article). In this way, the parameters of target encoder are an ensemble of the previous versions of online encoder [62] and updated in a slow pace. Compared with sharing parameters directly with online encoder, the generated target features change more smoothly in the iterations [22], and the online encoder can learn information from the past [63], which stabilizes the network training and improves the quality of features.

The original image (without masking) is input to the target encoder and the features at the first three scales are obtained as the target features T^1–T^3. For the output features of each 3D-FPN block (i.e., Q^1–Q^3), a $1 \times 1 \times 1$ convolution is used, respectively, to obtain the reconstructed features R^1–R^3. The loss function of MS-MFM is the sum of the $L1$ loss of reconstructed and target features at each scale

$$L_{\mathrm{MFM}} = \sum_{n=1}^{3} \frac{\sum M^n \odot |R^n - T^n|}{\sum M^n} \tag{11}$$

where M^n is the mask at each scale downsampled from the original mask M.

3) Total Loss: The proposed S^{2}HM^2 framework is trained with the combination of the spectral–spatial MIM and MS-MFM task, and the total loss can be expressed as

$$L = L_{\mathrm{MIM}} + \lambda L_{\mathrm{MFM}} \tag{12}$$

where λ is the weight hyperparameter that balances the contributions of these two tasks.

IV. RESULTS AND DISCUSSION

A. Dataset Description

To evaluate the performance of the proposed method, two large-scale hyperspectral datasets (i.e., WHU-OHS [3] and WHU-H^{2}SR [64]) and 14 other commonly used public hyperspectral datasets are selected for experiments. According to the main information of each dataset summarized in Table I, WHU-OHS and WHU-H^{2}SR (bolded in Table I) have much larger spatial coverage and data volume compared with other public hyperspectral datasets. The detailed description of these datasets is provided as follows.

1) WHU-OHS Large-Scale Hyperspectral Dataset: WHU-OHS [3] is a benchmark dataset for large-scale HSI classification composed of 42 Orbita Hyperspectral Satellites (OHS) images. The spatial resolution of the images is 10 m, and there are 32 spectral bands with wavelengths ranging from 466 to 940 nm. The images were acquired from more than 40 different regions of China, and the heterogeneity across different images is high. There are 7795 subimages with a size of 512×512 in the dataset, which are further cropped into 256×256 for experiments. In this research, we use the 34 images as source domain [3] for SSL pretraining and the eight images and corresponding labels (including 23 land cover classes) as target domain for fine-tuning and testing in the downstream HSI classification task. Specifically, there are 13 805 subimages for SSL pretraining, and 3329, 346, and 1115 subimages for training, validation, and testing of downstream task, respectively.

2) WHU-H^{2}SR Large-Scale High-Resolution and Hyperspectral Dataset: WHU-H^{2}SR [64] is a large-scale hyperspectral dataset with high spatial resolution. The images were acquired in the southern part of Shenyang, Liaoning, China,

TABLE II
SAMPLE DIVISION OF THE PUBLIC HYPERSPECTRAL DATASETS

Dataset	Training samples (pixels)	Testing samples (pixels)
Indian Pines	4959	4384
Pavia University	2427	33308
Pavia Center	48043	91133
Salinas	25859	20963
KSC	1701	3299
Botswana	1394	1824
DFC2013	2711	11406
DFC2018	237976	238193
Washington DC	759	18685
AeroRIT	3557139	2733574
WHU-Hi-HanChuan	1600	220736
WHU-Hi-HongHu	2200	352896
WHU-Hi-LongKou	900	176035
Matiwan	1948163	1728947

with a total area of 227.79 km^2. The spatial resolution is 1 m and there are 249 spectral bands with the spectral range between 391 and 984 nm. There are 1516, 253, and 762 subimages, including eight land cover classes with a size of 300 × 300 for training, validation, and testing, respectively. In the experiments, all images from the training set are used for SSL pretraining, and 20% of the images and labels in the training set are randomly chosen for fine-tuning. Please notice that since the WHU-H^{2}SR images cover only a single region, the landscapes across different subimages are more homogeneous compared with the WHU-OHS dataset.

3) Other Public Hyperspectral Datasets: The 14 public hyperspectral datasets are selected to evaluate the transferability of the proposed SSL framework (Table I). The official training and testing sets are adopted for Indian Pines,[1] Pavia University,[1] DFC2013,[2] Washington DC,[3] AeroRIT [65], and WHU-Hi [66]. For other datasets, we manually divide the samples into spatially disjoint training and testing sets. The number of training and testing samples for each dataset is shown in Table II.

B. Experimental Setting

The hyperparameters of the encoder are set with the reference to Swin Transformer-Tiny [49], where the number of feature maps C after 3-D patch embedding is 96. For WHU-OHS and WHU-H^{2}SR datasets, d is set to 4 and 16, respectively. According to the setting of FPN [57], the number of output feature maps $C^{'}$ of 3D-FPN is 256.

Experiments are divided into SSL pretraining and fine-tuning stages. For SSL pretraining, the mask ratio is 45% and 60% for WHU-OHS and WHU-H^{2}SR, respectively, and the size of each cube in the 3-D Mask is set to 4 × 32 × 32 and 16 × 32 × 32, which is equal to the patch size of the last block of the encoder [28] (see Section IV-E for the sensitivity analysis of the mask ratio and mask size). The momentum for updating the target encoder is 0.996 [22]. For simplicity, all parts of the loss function are treated equally, i.e., $\lambda_1 = \lambda_2 = \lambda = 1$. The network is pretrained for 100 and 200 epochs on WHU-OHS and WHU-H^{2}SR, and the initial learning rate is 0.0001 and 0.001, respectively, with a cosine decay schedule. The batch size is set to 8 and the AdamW optimizer is adopted.

After pretraining, the reconstruction head is replaced by a classification head (consisting of a 3 × 3 × 3 and a 1 × 1 × 1 convolution) [35], [68], with other parts of the network unchanged. The labels of HSI classification are used for fine-tuning the network. Given that the number of labels for each class is imbalanced in HSI classification [69], in the fine-tuning stage, weighted cross entropy [70] is chosen as a loss function for network training. The network is trained for 100 epochs with a batch size of 8 for both datasets. The learning rate is 0.00001 and 0.0001 for WHU-OHS and WHU-H^{2}SR, respectively.

OA, Kappa, mIoU, and IoU of each class are selected for accuracy evaluation of the downstream HSI classification task.

C. Comparison Experiments

1) Compared Methods: To demonstrate the effectiveness of the proposed S^{2}HM^2 on the downstream HSI classification task, the following two aspects of comparison experiments are designed.

1) *Comparison With Eight Supervised Algorithms:* Contextual CNN [71], SSRN [72], pResNet [39], HybridSN [73], A^{2}S^{2}K-ResNet [11], GMA-Net [12], GAHT [44], and HiT [14].
2) *Comparison With Random Initialization (i.e., Without Pretraining, Denoted as Random Init) and Six Self-Supervised Pretraining Algorithms:* SimCLR [20], MoCo v2 [74], BYOL [22], MOBY [75], MAEST [36], and SS-MTr [37]. For a fair comparison, the same encoder is adopted for pretraining and the same encoder–decoder architecture is used for fine-tuning. In addition, given the large domain difference between the source and target domains of WHU-OHS dataset [3], supervised pretraining using labels from source domain (denoted as Sup. init) is added to the comparison experiments.

2) Results on WHU-OHS Dataset: For WHU-OHS dataset, the accuracies of the proposed S^{2}HM^2 compared with supervised methods are shown in Table III, and a comparison with different pretraining algorithms is shown in Table IV, with the best and second-best accuracies highlighted in blue and bolded, respectively. In addition, the model parameters (Params), floating-point operations (FLOPs), and computational time for different algorithms are provided.

From Table III, we can see that the proposed S^{2}HM^2 achieves higher accuracies compared with the supervised algorithms, and the accuracies of most land cover classes are the best. As shown in Table IV, compared with random initialization, it is difficult for SSL methods based on contrastive learning (i.e., SimCLR, MoCo v2, BYOL, and MOBY) to stably improve the performance of downstream task. Specifically, MoCo v2 and MOBY obtain slightly better results, while SimCLR and BYOL have negative impacts. The MIM-based algorithms, i.e., MAEST, SS-MTr, and S^{2}HM^2, have reached better performance than random initialization,

[1] https://dase.grss-ieee.org/index.php
[2] https://hyperspectral.ee.uh.edu/?page_id=459
[3] https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

TABLE III
ACCURACIES OF THE PROPOSED S^2HM^2 COMPARED WITH SUPERVISED METHODS ON THE WHU-OHS DATASET

Class	Contextual CNN	SSRN	pResNet	HybridSN	A^2S^2K-ResNet	GMA-Net	GAHT	HiT	S^2HM^2
Paddy field	0.422	0.536	0.548	0.503	0.538	0.587	0.413	**0.643**	0.708
Dry farm	0.572	0.697	0.714	0.638	0.657	0.707	0.622	**0.758**	0.793
Woodland	0.724	0.787	0.732	0.763	0.769	0.806	0.761	**0.832**	0.847
Shrubbery	**0.076**	**0.076**	0.062	0.075	0.036	0.056	0.026	0.063	0.087
Sparse woodland	0.108	0.115	0.126	0.117	0.129	0.151	0.110	**0.217**	0.261
Other forest land	0.028	0.047	0.045	0.034	0.037	0.038	0.029	**0.094**	0.148
High-covered grassland	0.184	0.252	0.211	0.193	0.225	0.159	0.182	**0.264**	0.345
Medium-covered grassland	0.350	0.383	0.384	0.359	0.355	0.378	0.385	**0.427**	0.463
Low-covered grassland	0.548	0.560	0.601	0.562	0.563	0.585	0.572	0.634	**0.621**
River canal	0.748	0.759	0.793	0.771	0.798	0.798	0.772	**0.808**	0.817
Lake	0.926	**0.954**	0.928	0.936	0.935	0.942	0.935	0.963	**0.954**
Reservoir pond	0.317	0.351	0.306	0.297	0.288	0.372	0.218	**0.374**	0.469
Beach land	0.058	0.158	0.107	0.118	0.184	**0.228**	0.153	0.380	0.182
Shoal	0.117	0.232	0.248	0.166	0.242	0.284	0.200	**0.389**	0.439
Urban built-up	0.561	0.671	0.688	0.611	0.687	0.727	0.696	**0.772**	0.817
Rural settlement	0.325	0.405	0.399	0.350	0.374	0.448	0.396	**0.523**	0.593
Other construction land	0.224	0.297	0.326	0.259	0.333	0.399	0.293	**0.438**	0.478
Gobi	0.615	0.686	0.697	0.663	**0.701**	0.695	0.641	0.627	0.793
Saline-alkali soil	0.649	0.753	0.760	0.686	0.643	**0.794**	0.661	0.670	0.869
Marshland	0.277	0.534	0.374	0.473	0.478	0.436	0.410	0.669	**0.599**
Bare land	0.082	0.149	0.206	0.120	0.321	0.187	0.142	**0.496**	0.601
Bare rock	0.798	**0.860**	0.856	0.845	0.800	0.859	0.850	0.813	0.924
Ocean	0.942	0.941	0.978	0.953	0.951	0.973	**0.974**	0.960	0.966
OA	0.686	0.750	0.752	0.726	0.746	0.777	0.719	**0.815**	0.848
Kappa	0.658	0.727	0.729	0.700	0.722	0.755	0.693	**0.796**	0.832
mIoU	0.420	0.487	0.482	0.456	0.480	0.505	0.454	**0.557**	0.599
Params	0.30M	0.11M	1.10M	5.16M	0.11M	1.17M	0.73M	40.68M	31.22M
FLOPs	7.35M	13.89M	27.45M	273.93M	25.11M	60.49M	35.93M	594.11M	117.42G
Training time (s/epoch)	26145.77	20043.52	23278.27	43687.04	22562.29	28617.01	19859.03	47654.45	1033.15

** All the experiments are conducted on one NVIDIA GeForce RTX 4090 GPU.

as fine-grained representations from MIM are more suitable for such dense prediction task. Meanwhile, since MAEST and SS-MTr mask the HSI in only spectral or spatial dimension without the fully consideration on spectral–spatial redundancy, their classification accuracies are lower than the proposed S^2HM^2. Specifically, the OA of S^2HM^2 surpasses MAEST and SS-MTr by 3.6% and 3.0%, respectively, and reaching the highest among all compared methods. In addition, supervised pretraining using labels from the source domain also benefits the performance of HSI classification. However, the accuracy gain is smaller than the proposed method, indicating that the proposed S^2HM^2 learns more universal features. Figs. 4 and 5 show the visualization results for different methods, where the classification results of the proposed S^2HM^2 are more precise and closer to the real spatial distribution of objects. For example, in the second row of Fig. 4, the proposed method better extracts the paddy field, which is easily misclassified as sparse woodland by other compared methods. For the second row of Fig. 5, the main challenge lies in differentiating the grassland and dry farm with complex texture in the image. Compared with other methods, the proposed S^2HM^2 identifies the dry farm to a greater extent. In addition, the proposed method better depicts the shape of bare land in the bottom-left corner.

For the computational complexity, as shown in Table III, the FLOPs of the proposed method is higher than the supervised ones. The main reason lies in that the inputs of the supervised algorithms are overlapping patches generated in windows of small size (e.g., 7×7 for SSRN [72] and 9×9 for A^2S^2K-ResNet [11]). In contrast, the proposed method takes the whole image (i.e., subimages with the size of 256×256) as input. However, the patch-wise processing for the supervised algorithms exists huge computational cost and redundancy, resulting in much longer training time than the proposed "pixel-to-pixel" framework. With regard to SSL-based algorithms, it can be seen in Table IV that the proposed S^2HM^2 achieves the best performance without introducing too many extra parameters and FLOPs, and its efficiency is higher than that of contrastive learning-based methods and competitive to that of MIM-based ones.

3) Results on WHU-H^2SR Dataset: Quantitative results of the proposed S^2HM^2 compared with supervised algorithms and different pretraining algorithms on WHU-H^2SR dataset are presented in Tables V and VI, respectively. All algorithms use the same sample set (i.e., 20% of the labels randomly chosen from the training set) for network training or fine-tuning.

According to the results in Table V, for WHU-H^2SR dataset, the proposed S^2HM^2 achieves higher accuracies than the supervised algorithms, with the accuracy of most classes reaching the best (e.g., dry farmland, grassland, building, and water body) or second-best (e.g., paddy field, forest land, and greenhouse). In view of pretraining algorithms, as shown in Table VI, the performances of contrastive learning-based approaches are seriously degraded compared with random initialization. This phenomenon is partly due to the similar spatial landscapes of different images within WHU-H^2SR, as this dataset is acquired in a single region in Shenyang. In contrast, the MIM-based methods can benefit the downstream

TABLE IV
ACCURACIES OF DIFFERENT PRETRAINING ALGORITHMS ON THE WHU-OHS DATASET

Class	Random init	SimCLR	MoCo v2	BYOL	MOBY	MAEST	SS-MTr	Sup. init	S^2HM^2
Paddy field	0.635	0.613	0.641	0.620	0.656	0.647	0.650	**0.671**	0.708
Dry farm	0.712	0.710	0.730	0.730	**0.756**	0.716	0.736	0.745	0.793
Woodland	0.823	0.821	0.825	0.816	0.833	**0.834**	0.829	0.833	0.847
Shrubbery	**0.153**	0.071	0.124	0.095	0.125	0.102	0.146	0.156	0.087
Sparse woodland	0.212	**0.268**	0.272	0.244	0.242	0.193	0.214	0.215	0.261
Other forest land	0.089	0.075	0.094	0.068	0.098	0.095	0.096	**0.112**	0.148
High-covered grassland	0.258	0.236	0.228	0.238	0.263	0.241	0.266	**0.287**	0.345
Medium-covered grassland	0.412	0.396	0.427	0.434	**0.447**	0.430	0.441	0.424	0.463
Low-covered grassland	**0.598**	0.567	0.592	0.585	0.594	0.595	0.570	0.586	0.621
River canal	0.793	0.787	0.795	0.801	**0.806**	0.796	0.805	0.792	0.817
Lake	0.939	0.937	0.940	0.920	**0.947**	0.945	0.927	0.936	0.954
Reservoir pond	0.407	0.341	0.392	0.356	**0.437**	0.387	0.398	0.406	0.469
Beach land	0.221	0.136	0.157	0.147	0.132	0.376	**0.230**	0.225	0.182
Shoal	0.308	0.244	0.353	0.353	0.371	**0.401**	0.276	0.314	0.439
Urban built-up	0.776	0.752	0.779	0.781	0.756	0.749	**0.791**	0.774	0.817
Rural settlement	0.425	0.413	0.463	0.450	0.482	0.450	**0.492**	0.466	0.593
Other construction land	0.385	0.400	**0.428**	0.401	0.426	0.365	0.420	0.375	0.478
Gobi	0.712	0.674	0.719	0.704	0.691	0.746	**0.779**	0.771	0.793
Saline-alkali soil	0.775	**0.790**	0.756	0.720	0.773	0.680	0.491	0.676	0.869
Marshland	0.533	0.516	0.520	0.525	**0.591**	0.546	0.542	0.549	0.599
Bare land	**0.559**	0.309	0.460	0.350	0.513	0.329	0.508	0.528	0.601
Bare rock	0.884	0.863	0.873	0.840	0.816	0.876	0.898	**0.900**	0.924
Ocean	0.970	0.975	0.967	0.975	0.969	0.968	**0.974**	0.969	0.966
OA	0.808	0.793	0.809	0.800	0.815	0.812	0.818	**0.822**	0.848
Kappa	0.788	0.772	0.789	0.780	0.796	0.792	0.798	**0.803**	0.832
mIoU	0.547	0.517	0.545	0.528	0.553	0.542	0.543	**0.553**	0.599
Pre-training Params	N/A	29.43M	29.43M	33.91M	33.91M	30.75M	28.37M	31.22M	29.73M
Pre-training FLOPs	N/A	91.65G	91.65G	183.31G	183.31G	47.84G	46.12G	117.42G	107.00G
Pre-training time (s/epoch)	N/A	3252.66	3214.36	3251.79	3329.52	909.85	452.18	735.79	818.94

** The fine-tuning Params is 31.22M, FLOPs is 117.42G, and time is 214.21s/epoch for all the algorithms.
** All the experiments are conducted on one NVIDIA GeForce RTX 4090 GPU.

TABLE V
ACCURACIES OF THE PROPOSED S^2HM^2 COMPARED WITH SUPERVISED METHODS ON THE WHU-H^2SR DATASET

Class	Contextual CNN	SSRN	pResNet	HybridSN	A^2S^2K-ResNet	GMA-Net	GAHT	HiT	S^2HM^2
Paddy field	0.565	0.838	0.731	0.818	0.882	0.851	0.821	0.844	**0.874**
Dry farmland	0.429	0.631	0.642	0.591	0.723	**0.744**	0.683	0.719	0.773
Forest land	0.160	0.234	0.445	0.387	0.489	0.463	0.371	0.368	**0.474**
Grassland	0.023	0.141	0.144	0.101	0.187	**0.215**	0.174	0.178	0.222
Building	0.269	0.546	0.533	0.479	0.539	0.632	0.557	**0.569**	0.632
Highway	0.064	0.288	0.259	0.223	0.324	0.369	0.229	**0.338**	0.278
Greenhouse	0.025	0.496	0.403	0.249	0.503	0.565	0.524	0.477	**0.548**
Water body	0.368	0.276	0.381	0.450	**0.484**	0.396	0.389	0.393	0.490
OA	0.550	0.737	0.727	0.718	0.809	**0.815**	0.772	0.786	0.838
Kappa	0.380	0.639	0.629	0.614	0.733	**0.736**	0.679	0.701	0.763
mIoU	0.238	0.431	0.442	0.412	0.516	**0.529**	0.468	0.486	0.537
Params	1.27M	0.44M	1.13M	9.16M	0.45M	2.01M	0.98M	57.37M	31.33M
FLOPs	31.60M	119.90M	31.23M	3.12G	215.19M	686.76M	47.94M	1.51G	235.73G
Training time (s/epoch)	4936.12	6390.18	9862.25	36855.87	8640.88	41042.02	6214.60	13604.78	384.38

** All the experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs.

HSI classification task, and the proposed method significantly outperforms MAEST and SS-MTr. Figs. 6 and 7 illustrate the visualization results, where the proposed S^2HM^2 obtains the best classification performance among the compared methods. For instance, for the fourth row of Fig. 6, most of the compared methods fail to effectively extract the dry farmland (e.g., SSRN and GAHT) or the forest land (e.g., Contextual CNN and GMA-Net). In contrast, the proposed S^2HM^2 shows better separability between these two categories. In the first row of Fig. 7, the proposed method more accurately distinguishes between grassland, highway, and water body, and the paddy field in the bottom-right corner is more completely identified.

Furthermore, we select different ratio of labels (20%, 40%, 60%, 80%, and 100%) to fine-tune the model pretrained under the proposed S^2HM^2. As the number of labels increase, the change of classification accuracy is shown in Fig. 8, and the

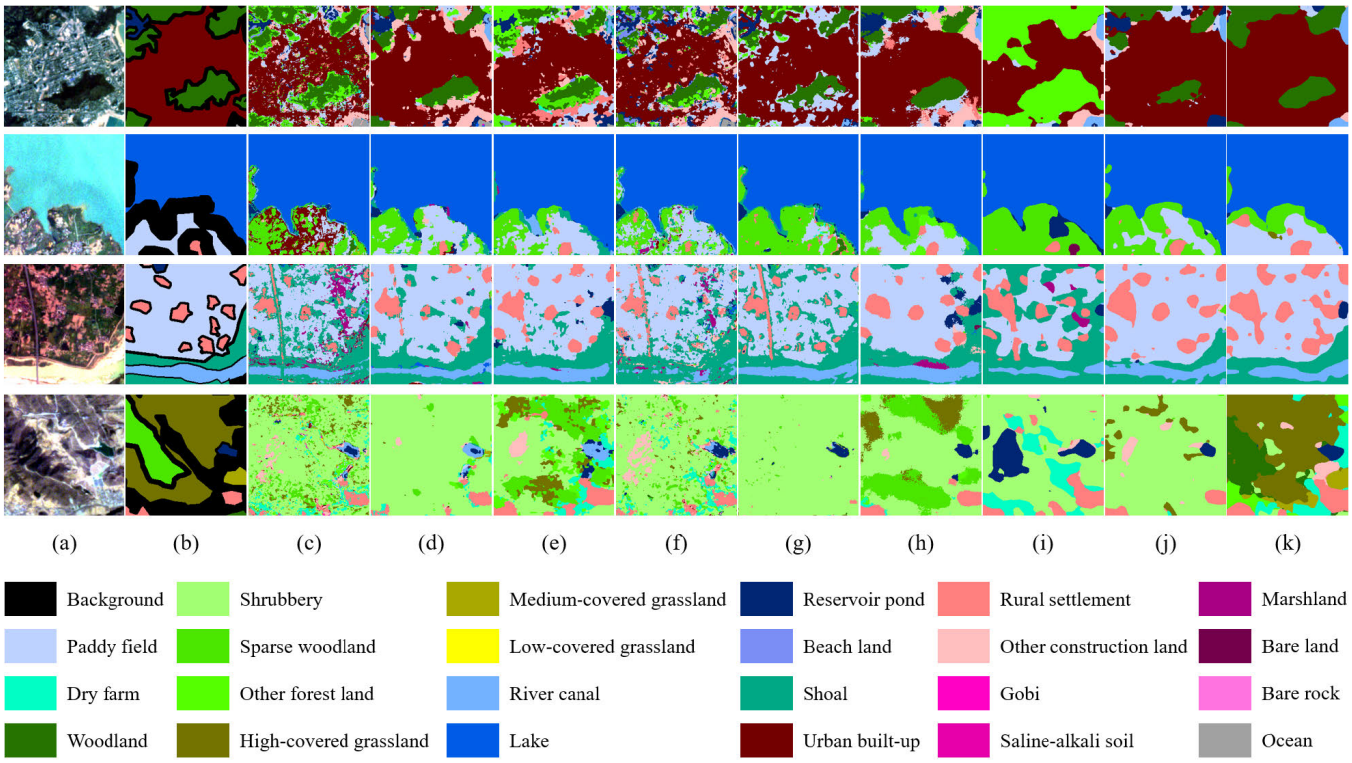


Fig. 4. Visualization of classification results of the proposed SSL framework compared with supervised algorithms on the WHU-OHS dataset. (a) Image (true-color compositions with R: 670 nm, G: 566 nm, and B: 480 nm). (b) Label. (c) ContextualCNN. (d) SSRN. (e) pResNet. (f) HybridSN. (g) A^2S^2K-ResNet. (h) GMA-Net. (i) GAHT. (j) HiT. (k) S^2HM^2.

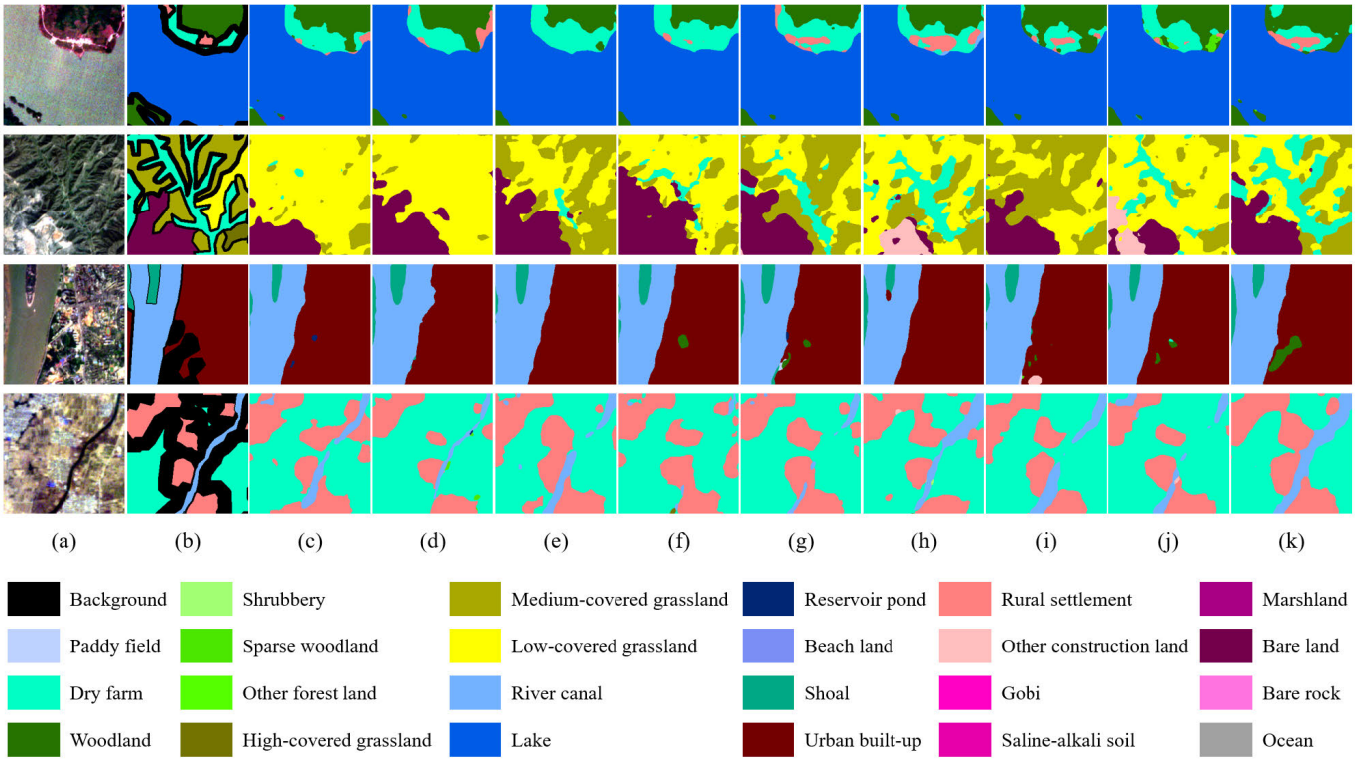


Fig. 5. Visualization of classification results of different pretraining algorithms on the WHU-OHS dataset. (a) Image. (b) Label. (c) Random init. (d) SimCLR. (e) MoCo v2. (f) BYOL. (g) MOBY. (h) MAEST. (i) SS-MTr. (j) Sup. init. (k) S^2HM^2.

proposed method can reach a relatively high accuracy with a small number of labels. For instance, when only 20% of the labels are used for fine-tuning, the accuracy of S^2HM^2 is comparative to training the network from random initialization

TABLE VI
ACCURACIES OF DIFFERENT PRETRAINING ALGORITHMS ON THE WHU-H^{2}SR DATASET

Class	Random init	SimCLR	MoCo v2	BYOL	MOBY	MAEST	SS-MTr	S^{2}HM^2
Paddy field	0.769	0.568	0.704	0.625	0.642	0.798	**0.810**	0.874
Dry farmland	0.645	0.506	0.559	0.475	0.515	0.665	**0.678**	0.773
Forest land	**0.389**	0.265	0.269	0.306	0.237	0.334	0.367	0.474
Grassland	0.168	0.114	0.121	0.132	0.107	**0.170**	0.134	0.222
Building	0.541	0.515	0.499	0.472	0.475	0.559	**0.562**	0.632
Highway	0.199	0.130	0.149	0.125	0.133	**0.232**	0.203	0.278
Greenhouse	**0.501**	0.189	0.195	0.158	0.085	0.473	0.480	0.548
Water body	**0.409**	0.314	0.319	0.293	0.258	0.379	0.338	0.490
OA	0.750	0.626	0.677	0.621	0.638	0.767	**0.778**	0.838
Kappa	0.649	0.473	0.552	0.485	0.495	0.664	**0.673**	0.763
mIoU	**0.453**	0.325	0.352	0.323	0.306	0.451	0.446	0.537
Pre-training Params	N/A	29.45M	29.45M	33.93M	33.93M	32.07M	28.38M	29.80M
Pre-training FLOPs	N/A	185.71G	185.71G	371.43G	371.43G	100.57G	92.71G	219.58G
Pre-training time (s/epoch)	N/A	1437.09	1427.81	1461.56	1803.91	513.48	234.77	245.28

** The fine-tuning Params is 31.33M, FLOPs is 235.73G, and time is 139.10s/epoch for all the algorithms.
** All the experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs.

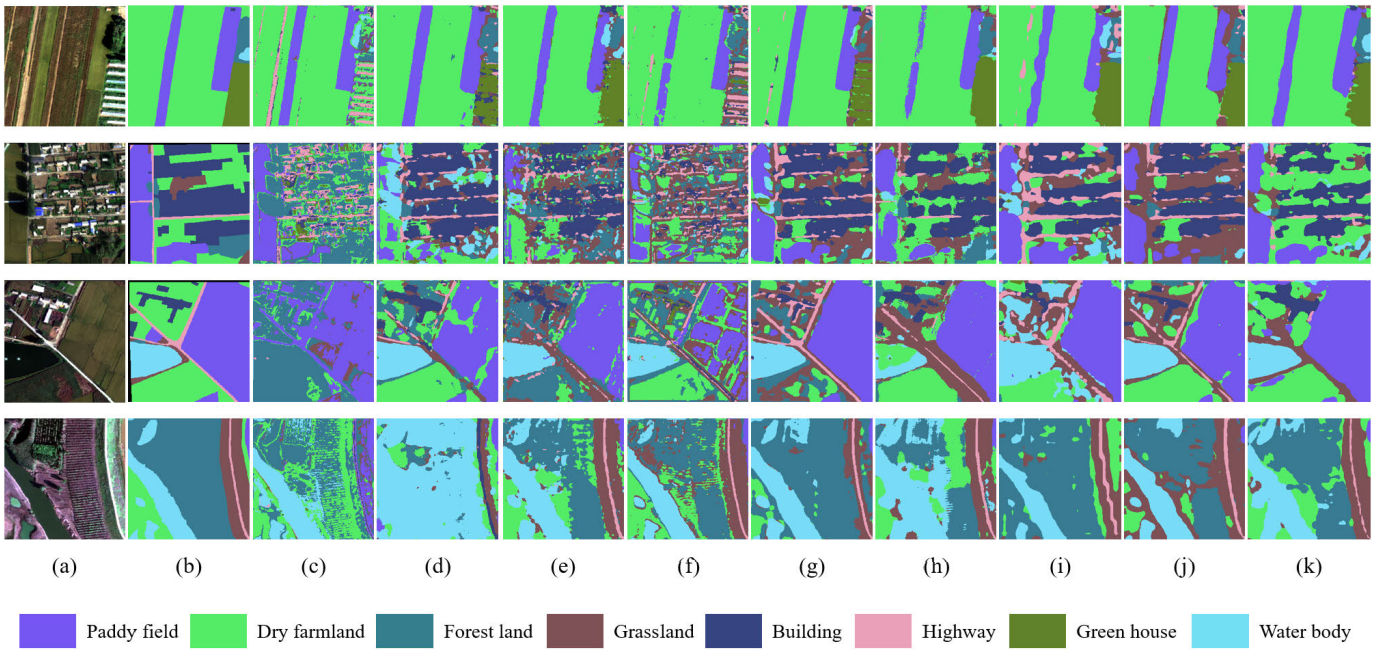


Fig. 6. Visualization of classification results of the proposed SSL framework compared with supervised algorithms on the WHU-H^{2}SR dataset. (a) Image (true-color compositions with R: 674 nm, G: 559 nm, and B: 474 nm). (b) Label. (c) ContextualCNN. (d) SSRN. (e) pResNet. (f) HybridSN (g) A^{2}S^{2}K-ResNet. (h) GMA-Net. (i) GAHT. (j) HiT. (k) S^{2}HM^2.

using all (100%) labels (with the OA of 83.8%). As the number of labels grows, both the accuracy of the proposed method and random initialization gradually increase, and the proposed S^{2}HM^2 consistently outperforms random initialization, with the highest OA reaching 86.8%. These results further verify the superiority of the proposed method when applied to the downstream HSI classification task.

D. Ablation Experiments

The main contributions of the proposed S^{2}HM^2 include the 3-D Mask strategy, the spatial–spectral consistency loss, the 3D-FPN, and the MS-MFM task. In this section, ablation experiments are conducted to investigate the effectiveness of these components.

1) Effectiveness of Spectral–Spatial 3-D Masking: Table VII shows the classification accuracies using the proposed 3-D Mask strategy compared with masking in only spectral (1-D Mask) or spatial (2-D Mask) dimension. The results indicate that the proposed 3-D Mask strategy helps the network to extract the spectral–spatial features more comprehensively. With the same spectral coverage (i.e., the wavelengths from 400 to 1000 nm), WHU-H^{2}SR image has much more spectral bands (i.e., 249) and finer spectral resolution than WHU-OHS image. Therefore, correlations between

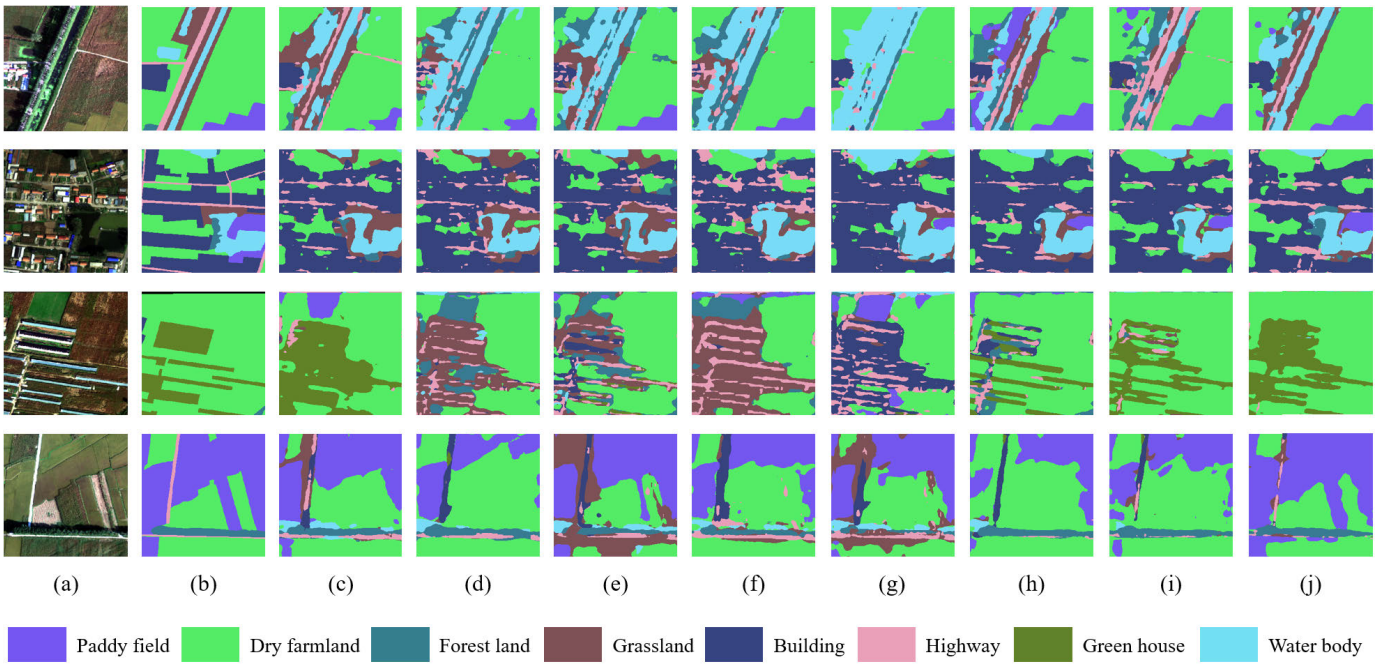


Fig. 7. Visualization of classification results of different pretraining algorithms on the WHU-H²SR dataset. (a) Image. (b) Label. (c) Random init. (d) SimCLR. (e) MoCo v2. (f) BYOL. (g) MOBY. (h) MAEST. (i) SS-MTr. (j) S²HM².

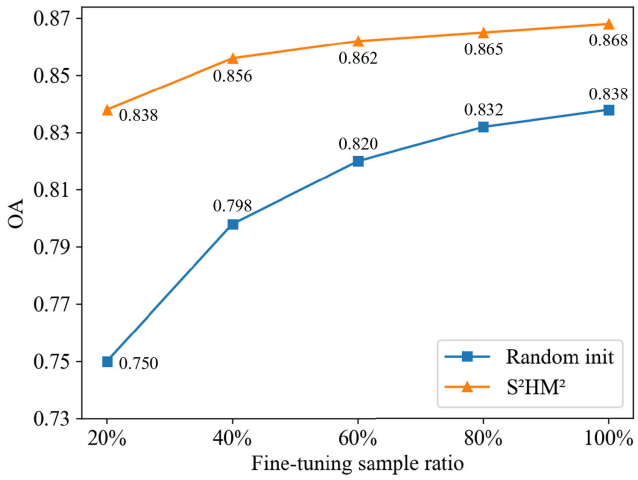


Fig. 8. Comparison of classification accuracy with different sample ratios on the WHU-H²SR dataset.

TABLE VII
ABLATION RESULTS FOR DIFFERENT MASKING STRATEGIES

Masking stretagy	WHU-OHS			WHU-H²SR		
	OA	Kappa	mIoU	OA	Kappa	mIoU
1D Mask	0.835	0.817	0.570	0.770	0.683	0.471
2D Mask	0.842	0.826	0.585	0.803	0.715	0.494
3D Mask	0.848	0.832	0.599	0.838	0.763	0.537

TABLE VIII
ABLATION RESULTS FOR THE OPTIMIZATION OBJECTIVE OF MIM

MIM Loss	WHU-OHS			WHU-H²SR		
	OA	Kappa	mIoU	OA	Kappa	mIoU
L1	0.845	0.829	0.586	0.827	0.750	0.525
L1+SAM	0.846	0.830	0.591	0.836	0.761	0.531
L1+SSIM	0.845	0.829	0.592	0.831	0.756	0.533
Proposed	0.848	0.832	0.599	0.838	0.763	0.537

different bands of WHU-H²SR image are much higher, and the low-rank and sparse properties are more significant. In this scenario, it is more significant to adopt the 3-D Mask strategy, where the masked images can be reconstructed with the cue in both spectral and spatial dimensions, to learn the intrinsic features of the images. As a result, the proposed 3-D Mask strategy boost the performance on WHU-H²SR dataset more significantly compared with that on WHU-OHS dataset.

2) Effectiveness of Spectral–Spatial Consistency Loss: The classification accuracies with traditional $L1$ loss, the combination of $L1$ with SAM or SSIM loss (denoted as $L1$+SAM and $L1$+SSIM, respectively) and the proposed spectral–spatial consistency loss (consisting of $L1$, SAM, and SSIM losses) are reported in Table VIII. The results show that both SAM and SSIM losses have positive effects on the feature learning process, and the proposed loss achieves the best performance with an overall improvement of 0.2% and 1.1% in OA on WHU-OHS and WHU-H²SR, respectively. As visually presented in Fig. 9, the reconstruction images optimized by the proposed loss can express more spatial details (see yellow boxes for instance) as well as is closer to the original spectral curves [Fig. 9(e)]. To sum up, the proposed spectral–spatial consistency loss achieves both better reconstruction results and downstream accuracies.

3) Effectiveness of 3D-FPN and MS-MFM: To verify the effectiveness of the proposed 3D-FPN and MS-MFM, SSL pretraining without using 3D-FPN, i.e., the output features of the last encoder layer are directly input into a linear layer for reconstruction [28], is taken as the baseline, whose

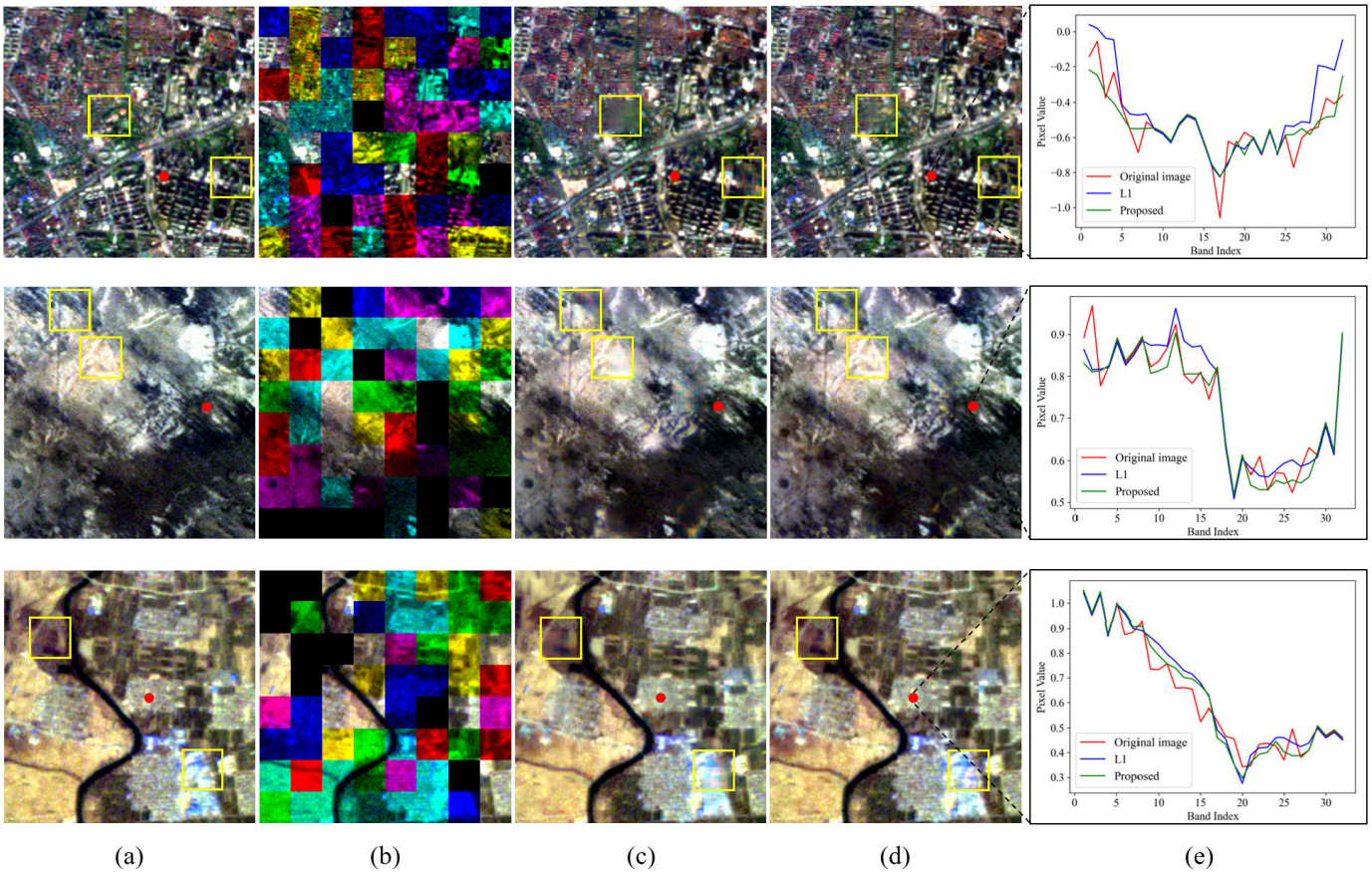


Fig. 9. Visualization results of reconstruction with different loss functions of MIM. (a) Original image. (b) Masked image (visualized by setting the pixel value of the masked patches to 0). (c) Reconstruction result using $L1$ loss. (d) Reconstruction result using the proposed spectral–spatial consistency loss. (e) Comparison of spectral curves (pixels marked by red points).

TABLE IX
ABLATION RESULTS FOR THE 3D-FPN AND MS-MFM

3D-FPN	MS-MFM	WHU-OHS			WHU-H^{2}SR		
		OA	Kappa	mIoU	OA	Kappa	mIoU
×	×	0.822	0.802	0.545	0.759	0.648	0.424
✓	×	0.841	0.824	0.584	0.812	0.726	0.506
✓	Weight-sharing	0.844	0.827	0.591	0.774	0.685	0.478
✓	Momentum	0.848	0.832	0.599	0.838	0.763	0.537

performance is listed in the first line in Table IX. Also, the momentum update for the target encoder in MS-MFM is compared with the simple weight-sharing strategy (i.e., the third line in Table IX). As shown in Table IX, when the network is pretrained without 3D-FPN, the OAs on WHU-OHS and WHU-H^{2}SR are 82.2% and 75.9%, respectively. After the proposed 3D-FPN is adopted, the accuracies in both datasets are significantly raised, with an improvement of 1.9% and 5.3%. It can be indicated that the proposed 3D-FPN enables the network to fully exploit the features at each scale, thus benefiting the downstream HSI classification task. The introduction of MS-MFM task further facilitates the multiscale feature extraction, leading to an increment of 0.7% and 2.6% in OA for the two datasets. Compared to sharing weights with online encoder, the momentum update process generates the reconstruction targets with higher quality. In particular, for WHU-H^{2}SR dataset with a relatively smaller amount of training data, it is more necessary to stabilize the training process with the momentum update.

E. Sensitivity Analysis on Mask Ratio and Mask Size

For the MIM-based works, mask ratio and mask size can influence the performance of both feature learning and downstream task [27], [28]. Therefore, in this section, sensitivity analysis on mask ratio and mask size is conducted. The classification accuracies with different mask ratios, and different mask sizes in spatial or spectral dimension, are shown in Fig. 10.

As the mask ratio and mask size increase, the overall performance on both datasets gradually increases at first. However, too high mask size or mask ratio can have negative impacts on the accuracy, which is mainly due to the large amount of information loss preventing the effective feature learning. For WHU-OHS dataset, as seen in Fig. 10(a)–10(c), the proposed method consistently reaches a relatively high classification accuracy with different mask ratios [with OA ranging from 84.3% to 84.8% in Fig. 10(a)] and mask sizes [with OA ranging from 83.8% to 84.8% for different spatial mask sizes in Fig. 10(b) and from 84.3% to 84.8% for different spectral mask sizes in Fig. 10(c)]. One possible reason is that the spatial resolution of the OHS images is relatively low (i.e., 10 m). Even though the mask ratio or mask size is small, the masked

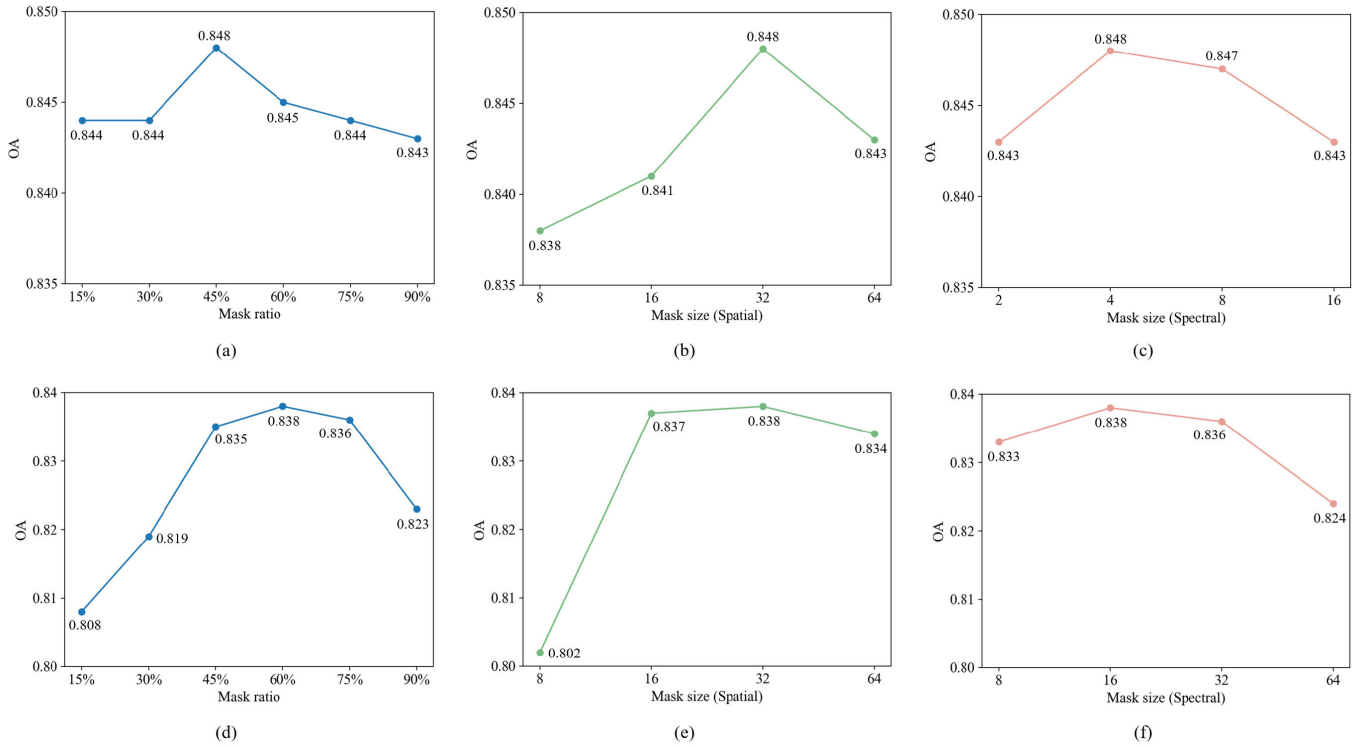


Fig. 10. Classification accuracies with different mask ratios and mask sizes. (a)–(c) WHU-OHS dataset. (d)–(f) WHU-H^{2}SR dataset. (a) and (d) Accuracies with different mask ratios, where the mask sizes are fixed to 4 × 32 × 32 and 16 × 32 × 32 for WHU-OHS and WHU-H^{2}SR, respectively. (b) and (e) Accuracies with different mask sizes in spatial dimension, where the mask sizes in spectral dimension are fixed to 4 and 16, and the mask ratios are fixed to 45% and 60% for WHU-OHS and WHU-H^{2}SR. (c) and (f) Accuracies with different mask sizes in spectral dimension, where the mask size in spatial dimension is fixed to 32 for both datasets.

patches are distant enough from neighboring visible patches, enforcing the network to learn long-range connections [28]. In contrast, for WHU-H^{2}SR dataset, the accuracy is seriously degraded when the mask ratio or mask size is too low or too high. The possible reasons can be summarized as follows.

1) In view of spatial information mining, the resolution of WHU-H^{2}SR image is high (i.e., 1 m). Therefore, when the mask ratio [Fig. 10(d)] or the spatial mask size [Fig. 10(e)] is too low, the masked parts can easily be recovered using the neighboring pixels, which hinders the high-level understanding of the images [27].
2) In view of spectral information representation, the large number of bands makes it difficult for reconstruction of the original spectral curves when the mask ratio or the spectral mask size [Fig. 10(f)] is too high.

Nonetheless, the proposed S^{2}HM^2 performs well on WHU-H^{2}SR dataset with a wide range of mask ratios (45%–75%) and mask sizes (16–64 in spatial dimension and 8–32 in spectral dimension). According to the experimental results in Fig. 10, the optimal mask ratio on WHU-OHS and WHU-H^{2}SR datasets is 45% and 60%, and the optimal mask size is 4 × 32 × 32 (i.e., four in spectral dimension and 32 in spatial dimension) and 16 × 32 × 32, respectively.

F. Effects of the Amount of Data for Self-Supervised Pretraining

As mentioned before, SSL can leverage large amount of unlabeled data for feature learning. In this section, the effects of the data volume for SSL pretraining are investigated. Taking WHU-OHS dataset as an example, we randomly select the subset of the pretraining data with the ratio of 20%, 40%, 60%, and 80% to train the network, and the accuracies on the downstream task are presented in Fig. 11. Compared with random initialization, the proposed method can achieve an increment of approximately 3% in OA even when only 20% of the data are used for SSL pretraining. As the amount of pretraining data increases, the classification accuracy is gradually boosted. It can be predicted that the performance

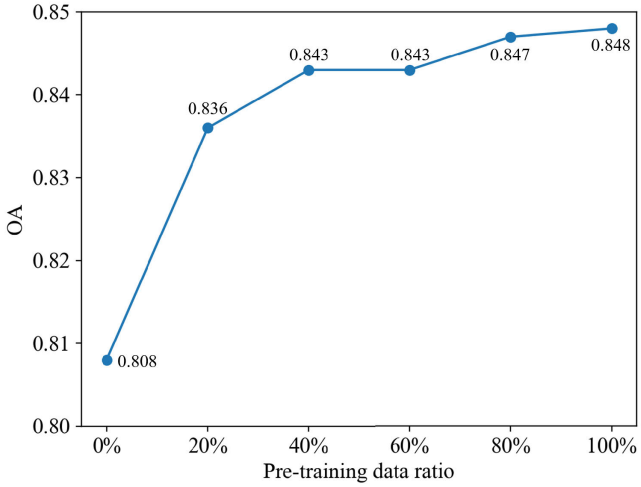


Fig. 11. Classification accuracies with different amount of pretraining data (0% represents random initialization without pretraining).

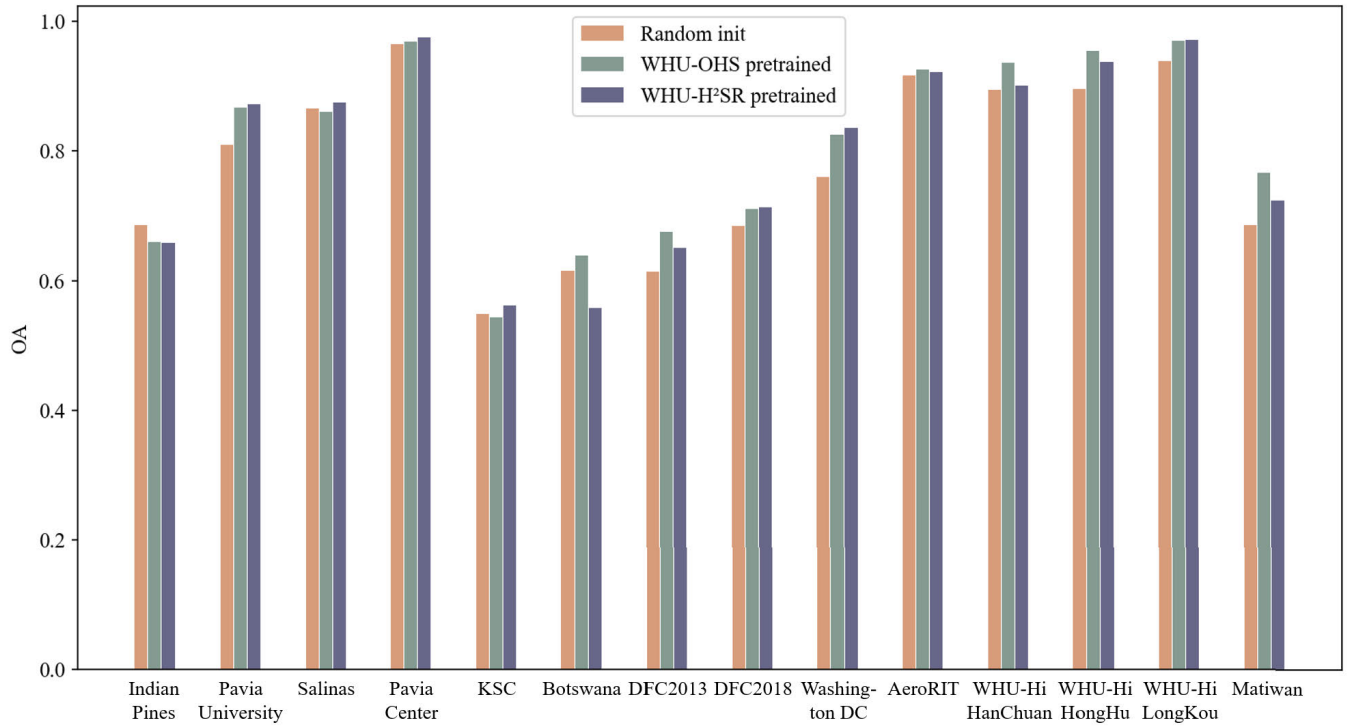


Fig. 12. Results of transferring to public hyperspectral datasets.

TABLE X
RESULTS OF TRANSFER LEARNING BETWEEN THE WHU-OHS AND WHU-H^2SR DATASETS

Pre-training	Fine-tuning					
	WHU-OHS			WHU-H^2SR		
	OA	Kappa	mIoU	OA	Kappa	mIoU
Random init	0.808	0.788	0.547	0.750	0.649	0.453
WHU-OHS	0.848	0.832	0.599	0.844	0.774	0.560
WHU-H^2SR	0.814	0.795	0.547	0.838	0.763	0.537

of the proposed S^2HM^2 will be further improved if the larger volume of data is introduced for pretraining in the future.

G. Transferability of the Proposed Method

In this section, experiments of transfer learning are conducted, to analyze the generalization ability of the proposed S^2HM^2, given heterogeneous hyperspectral datasets for pretext task and downstream task. Experiments are divided into two parts: transferring between WHU-OHS and WHU-H^2SR datasets, and transferring to other 14 public hyperspectral datasets (introduced in Section IV-A).

1) Transferring Between WHU-OHS and WHU-H^2SR Datasets: In this experiment, images and labels in the WHU-OHS dataset are utilized to fine-tune the model pretrained on WHU-H^2SR dataset and vice versa.

The results in Table X demonstrate the transferability of the proposed method. Compared with random initialization, the accuracies of downstream task are improved when the model is pretrained on either dataset. Specifically, after fine-tuning the model pretrained on WHU-H^2SR dataset, the OA on WHU-OHS dataset is increased by 0.6%. When the model pretrained on WHU-OHS dataset is transferred to WHU-H^2SR, the OA is significantly boosted by 9.4% and even surpasses pretraining and fine-tuning on WHU-H^2SR itself by 0.6%. It can be inferred that the data volume for SSL pretraining has vital impacts on the model transferability. By virtue of larger size and diversity of the WHU-OHS images, the model pretrained on WHU-OHS dataset achieves better accuracies than that pretrained on WHU-H^2SR dataset.

2) Transferring to Other Public Hyperspectral Datasets: Given the proposed S^2HM^2, the models pretrained on the large-scale datasets are transferred to the 14 public hyperspectral datasets. The training samples on each dataset are used to fine-tune the model.

As illustrated in Fig. 12, the models pretrained on large-scale hyperspectral datasets improve the downstream performance on most target datasets. Specifically, when the WHU-OHS dataset is used for pretraining, accuracies on 11 datasets (except for Indian Pines, Salinas, and KSC) outperform random initialization. When the WHU-H^2SR dataset is used for pretraining, accuracies on 12 datasets (except for Indian Pines and Botswana) are higher than random initialization. Therefore, the proposed S^2HM^2 exhibits promising transferability.

Taking a closer look at the characteristics and results of each dataset, the following factors could influence the performance of transfer learning.

a) Spectral range: The wavelengths of spectral bands for the two pretraining datasets are within the range of 400–1000 nm, including visible and near-infrared (VNIR) regions of the electromagnetic spectrum. Among the public hyperspectral datasets, Indian Pines, Salinas, KSC, and Botswana cover the wavelengths from 400 to 2500 nm, which includes VNIR and short-wave infrared (SWIR) regions and is the spectral coverage for most of the airborne or spaceborne

hyperspectral sensors [76]. The difference in spectral range can partly explain the reason that the pretrained models may not work well on these datasets. Therefore, it is suggested to construct the hyperspectral dataset with the wavelengths ranging from 400 to 2500 nm for pretraining, to increase the robustness of models.

b) Spatial resolution: For hyperspectral datasets with the spatial resolution close to WHU-H^2SR (i.e., Pavia University, Pavia Center, DFC2018, and Washington DC with the spatial resolution close to 1 m), the model pretrained on WHU-H^2SR dataset outperforms that pretrained on WHU-OHS dataset. Therefore, using data with similar spatial resolution for pretraining is beneficial for the performance of downstream task.

c) Data size and diversity: The OHS images have 32 spectral bands with a spatial resolution of 10 m, which are extremely different from these hyperspectral datasets. However, the model pretrained on WHU-OHS dataset significantly improves the downstream performance on most of the datasets, with higher classification accuracies than WHU-H^2SR pretraining on about half of them. It can be indicated that high data volume and data heterogeneity (i.e., 34 images with different acquisition time and location for pretraining) enable the model pretrained on WHU-OHS dataset to be more adaptive to the domain differences, and therefore, the amount and diversity of the pretraining data are important factors to be considered in transfer learning.

V. Conclusion

In this article, an SSL framework based on S^2HM^2 has been proposed for large-scale HSI classification. The proposed spectral–spatial 3-D masking strategy and spectral–spatial consistency loss jointly construct a better MIM task, facilitating a more comprehensive understanding on the spectral–spatial characteristics of HSI. The proposed hierarchical 3D-FPN decoder can fully exploit the multiscale features that are advantageous for HSI classification. In addition, the proposed MS-MFM task reconstructs the encoding features at each scale from the corresponding decoding features, further enhancing the multiscale feature learning. The experimental results on the two large-scale hyperspectral datasets, i.e., WHU-OHS and WHU-H^2SR, indicate that the proposed S^2HM^2 captures effective representations from large amount of unlabeled data and achieves the classification accuracies higher than existing supervised and SSL algorithms. Moreover, the proposed method shows promising transferability on a variety of public hyperspectral datasets. For future work, SSL with larger volume of HSI pretraining data and prompt learning to align more downstream tasks (e.g., spectral unmixing, anomaly detection, and crop yield estimation) are listed on the agenda, with the aim of designing a foundation model in the field of hyperspectral remote sensing.

Acknowledgment

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

References

[1] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, no. 1, pp. 110–122, Sep. 2009.

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

[3] J. Li, X. Huang, and L. Tu, "WHU-OHS: A benchmark dataset for large-scale hersepctral image classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, Sep. 2022, Art. no. 103022.

[4] R. Jain and R. U. Sharma, "Airborne hyperspectral data for mineral mapping in Southeastern Rajasthan, India," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 81, pp. 137–145, Sep. 2019.

[5] I. Aneece, D. Foley, P. Thenkabail, A. Oliphant, and P. Teluguntla, "New generation hyperspectral data from DESIS compared to high spatial resolution PlanetScope data for crop type classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7846–7858, 2022.

[6] Y. Gao et al., "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512615.

[7] F. Chen, H. Jiang, T. Van De Voorde, S. Lu, W. Xu, and Y. Zhou, "Land cover mapping in urban environments using hyperspectral APEX data: A study case in Baden, Switzerland," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 71, pp. 70–82, Sep. 2018.

[8] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[9] 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.

[10] 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.

[11] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

[12] T. Lu, M. Liu, W. Fu, and X. Kang, "Grouped multi-attention network for hyperspectral image spectral–spatial classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507912.

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

[14] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[15] 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.

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

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

[18] 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, 2022, Art. no. 5618014.

[19] 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, 2022, Art. no. 5411415.

[20] 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. 1575–1585.

[21] 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. 9726–9735.

[22] J. B. Grill, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–14.

[23] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[24] 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.

[25] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[26] C. Tao et al., "Siamese image modeling for self-supervised vision representation learning," 2022, *arXiv:2206.01204*.

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

[28] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9643–9653.

[29] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," 2022, *arXiv:2208.00173*.

[30] X. Chen et al., "Context autoencoder for self-supervised representation learning," 2022, *arXiv:2202.03026*.

[31] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315.

[32] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5612822.

[33] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "CMID: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607817.

[34] Q. He et al., "AST: Adaptive self-supervised transformer for optical remote sensing representation," *ISPRS J. Photogramm. Remote Sens.*, vol. 200, pp. 41–54, Jun. 2023.

[35] Y. Tian et al., "Fast-iTPN: Integrally pre-trained transformer pyramid network with token migration," 2022, *arXiv:2211.12735*.

[36] D. Ibañez, 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, 2022, Art. no. 5542614.

[37] 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, 2023, Art. no. 5508718.

[38] 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.

[39] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for Spectral–Spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[40] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.

[41] A. Vaswani, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 5999–6009.

[42] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

[43] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[44] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

[45] R. Ji, K. Tan, X. Wang, C. Pan, and L. Xin, "PASSNet: A spatial–spectral feature extraction network with patch attention module for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[46] 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, 2023, Art. no. 4401118.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[48] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[49] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[50] 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.

[51] 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, 2022, Art. no. 5526714.

[52] P. Khosla, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–12.

[53] Q. Liu, J. Peng, G. Zhang, W. Sun, and Q. Du, "Deep contrastive learning network for small-sample hyperspectral image classification," *J. Remote Sens.*, vol. 3, p. 25, Mar. 2023.

[54] 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.

[55] J. Li, X. Li, Z. Cao, and L. Zhao, "ROBYOL: Random-occlusion-based BYOL for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[56] 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, 2023, Art. no. 5531614.

[57] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

[58] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[59] S. Ertürk, "Fuzzy fusion of change vector analysis and spectral angle mapper for hyperspectral change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 5045–5048.

[60] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[61] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 157–176, Jun. 2022.

[62] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1196–1205.

[63] T. Pham, C. Zhang, A. Niu, K. Zhang, and C. D. Yoo, "On the pros and cons of momentum encoder in self-supervised visual representation learning," 2022, *arXiv:2208.05744*.

[64] B. Yu, J. Li, and X. Huang, "STSNet: A cross-spatial resolution multimodal remote sensing deep fusion network for high resolution land-cover classification," unpublished.

[65] A. Rangnekar, N. Mokashi, E. J. Ientilucci, C. Kanan, and M. J. Hoffman, "AeroRIT: A new scene for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8116–8124, Nov. 2020.

[66] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "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.

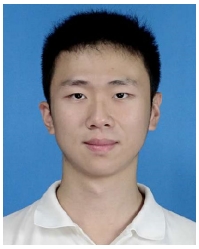
[67] Y. Cen et al., "Aerial hyperspectral remote sensing classification dataset of Xiongan New Area (Matiwan Village)," *J. Remote Sens.*, vol. 24, no. 11, pp. 1299–1306, 2020.

[68] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," 2018, *arXiv:1807.10221*.

[69] Y. Su, X. Li, J. Yao, C. Dong, and Y. Wang, "A spectral–spatial feature rotation-based ensemble method for imbalanced hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515918.

[70] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10795–10816, Sep. 2023.

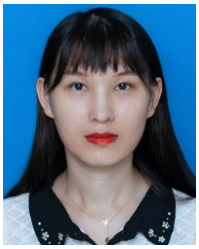
[71] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
[72] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[73] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
[74] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.
[75] Z. Xie et al., "Self-supervised learning with Swin transformers," 2021, *arXiv:2105.04553*.
[76] P. Ghamisi et al., "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.



Lilin Tu received the B.S. degree from Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering.

His research interests include hyperspectral image classification, semantic segmentation, change detection, and deep learning.

Mr. Tu won the second place in the 2021 IEEE GRSS Data Fusion Contest—Track MSD: Multitemporal Semantic Change Detection.



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 currently an Associate Professor with the School of Remote Sensing and Information Engineering and Hubei Luojia Laboratory, 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 and pattern recognition, and remote sensing images.

Dr. Li is a Young Editorial Board Member of *Geo-Spatial information Science* (GSIS) and a Guest Editor of *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, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, Remote Sensing of Environment (RSE), and ISPRS *Journal of Photogrammetry and Remote Sensing* (ISPRS-J).

Xin Huang (Senior Member, IEEE) received 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 2009.

He is currently a Full Professor with Wuhan University, where he teaches remote sensing, image interpretation, and so on. He is also the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University. He has been supported by the National Program for Support of Top-Notch Young Professionals in 2017, 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) articles) in the international journals. His research interests include remote sensing image processing methods and applications.

Dr. Huang has been an Editorial Board Member of 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. He was the winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was an Associate Editor of *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2014 to 2020, and 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 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2022.

Jianya Gong received the Ph.D. degree in photogrammetry and remote sensing from Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1992.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He is also an Academician with Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image processing, spatial data infrastructure, and geospatial data sharing and interoperability.

Xing Xie is currently with the School of Environmental and Mapping Engineering, Suzhou University, Suzhou, China.

His research interests include Gabor filters, image classification, image fusion, image segmentation, image texture, and parameter estimation.

Leiguang Wang received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, Hubei, China, in 2009.

From 2014 to 2015, he was a Post-Doctoral Researcher with the University of New Brunswick, Fredericton, NB, Canada. He is currently a Professor with Southwest Forestry University, Kunming, China, where he has been the Vice Dean of the Institute of Big Data and Artificial Intelligence since 2018. He is the author of more than 50 articles. His research interests include remote sensing image fusion, semantic segmentation, and application in forestry.