A Universal Representation Mechanism for Multisource Hyperspectral Remote Sensing Image

Weili Kong, *Graduate Student Member, IEEE*, Baisen Liu, Xiaojun Bi, and Yihan He

***Abstract*— Deep learning-based processing of hyperspectral remote sensing images (HSIs) has emerged as a research hotspot. In recent years, the delivery of numerous novel HSI acquisition platforms has resulted in an exponential growth in the number of available HSI datasets from various sources. However, due to variations among hyperspectral sensors, the acquired data frequently consists of various spectral dimensions. This results in a challenge since standard deep learning approaches often require a different model for each HSI source, which impedes the construction of fundamental models for HSIs. To address this issue, we propose a unified representation mechanism for multisource HSIs that can transform spectra from numerous dimensions to a shared representation space, yielding a scalable pretraining model. Compared to existing methods, our strategy has the following advantages: 1) compatibility with HSIs of arbitrary spectral dimensions, ranges, and resolutions; 2) fully leveraging existing multisource HSIs; 3) spontaneously capturing spectral features via self-supervised learning; and 4) pretrained on large-scale multisource HSI datasets and a considerable enhancement in classification accuracy. In three reconstruction test sets, the PSNRs are 27.07, 22.27, and 29.35 dB, and the SSIMs are 0.93, 0.82, and 0.88, respectively. Compared with the randomly initialized model, there are 8.28% and 1.4% improvements on the Indian Pines dataset and Pavia University dataset respectively.**

***Index Terms*— Deep learning, foundational model, multisource hyperspectral remote sensing images (HSIs), self-supervised learning, unified representation.**

I. Introduction

HYPERSPECTRAL remote sensing technology offers a powerful tool for earth science research by offering detailed spectral information across a wide range of wavelengths [1]. This characteristic enables researchers to conduct meticulous analysis of the observed objects [2]. However, precisely for this property, such data exhibit extremely high dimensionality. This will lead to the "curse of dimensionality," and heavy computational costs [3]. Representation learning stands as the optimal means to address this issue [4]. It can effectively diminish the redundant data, push out the crucial information, and attain more informative representations so enabling the model to perform classification and recognition tasks more accurately. In the early stages, scholars employed machine learning to learn the primary components of HSI data for representation. Typical methods include principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA), and so on [5]. However, such linear methods still have certain limitations. They struggle to capture the nonlinear variations within the data [6]. Therefore, scholars shifted their focus to methods grounded in deep learning [7], [8]. This category of methods can effectively capture the nonlinear structures within the data, and the most representative approach is autoencoder (AE) [9]. Such as Zhou et al. [10] designed a feature extraction technique utilizing semi-supervised stacked autoencoders (semi-SAE). In our previous work [11], we designed a pretraining network based on masked AEs to extract universal spatial–spectral features of HSI through self-supervised learning. Furthermore, some research has focused on representing hyperspectral data by a combination of mechanism and deep learning. For example, Kang et al. [12] proposed a 2-D spectral representation, which converts spectral information into images. It effectively improves the classification accuracy of spectral signals and provides a novel spectral visualization method.

Despite the significant improvements achieved by the aforementioned methodologies, the initialized models can not adapt to spectral data of different dimensions due to the constraints of matrix operations. Nowadays, an increasing number of HSI capture missions are being deployed, such as MODIS, HypSEO, DESIS, Gaofen-5, EnMap, HyspIRI, and so on [13]. These HSI acquisition platforms furnish an exceptionally rich wellspring of data, which could promote the advancement of deep learning in this field. However, unlike ordinary images that typically have only RGB channels or a single grayscale channel, multisource hyperspectral remote sensing images (HSIs) often possess varying spectral channels according to different observation requirements and sensor parameters [14]. This discrepancy complicates the processing of multisource HSI using deep learning approaches, as models often require different structures and training from scratch [15]. It implies the abundance of multisource HSIs cannot be directly beneficial to deep learning models, indicating an enormous gap in developing fundamental models specific to multisource HSIs. To fill this gap, we construct a unified representation

Manuscript received 27 April 2024; revised 4 June 2024; accepted 19 June 2024. Date of publication 24 June 2024; date of current version 4 July 2024. This work was supported in part by the Natural Science Foundation of Heilongjiang Province for Key Projects, China, under Grant ZD2021F004; and in part by the 2023 Heilongjiang Province Art and Science Planning Key Projects under Grant 2023A010. *(Corresponding author: Baisen Liu.)*

Weili Kong and Yihan He are with the School of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: kkweil@hrbeu.edu.cn; heyihan@hrbeu.edu.cn).

Baisen Liu is with the School of Electronic and Information Engineering, Heilongjiang Institute of Technology, Harbin 150050, China, and also with the School of Information and Communication Engineering, Harbin Engineering University, Harbin 150050, China (e-mail: spedliu@126.com).

Xiaojun Bi is with the School of Information Engineering and the Key Laboratory of Ethnic Language Intelligent and Analysis Security Governance of MOE, Minzu University of China, Beijing 100081, China (e-mail: bixiaojun@hrbeu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2024.3418856

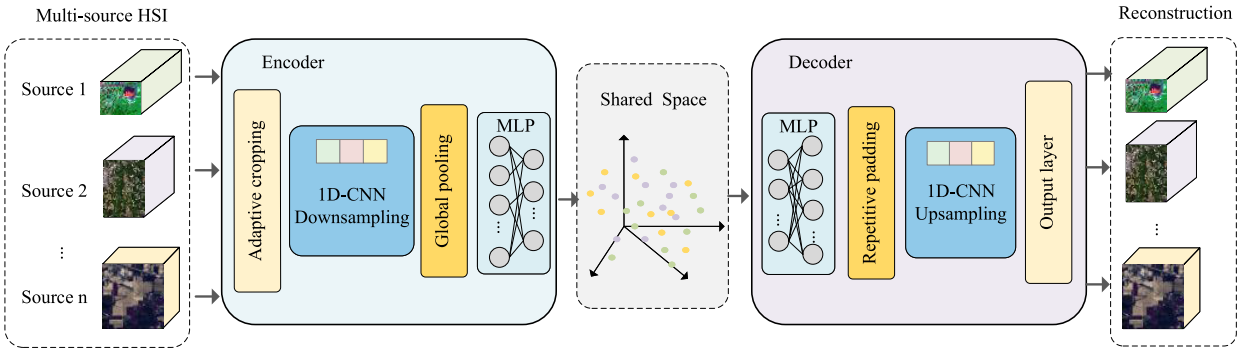

Fig. 1. Overall architecture of shared encoder. In the encoder section, we first apply adaptive cropping to multisource HSIs, then employ a 1D-CNN downsampling module for feature extraction, and finally perform global pooling to extract key factors and harmonize dimensional differences across the datasets. In the decoder section, a repetitive padding layer and a 1D CNN upsampling model are used to perform the reverse operations corresponding to the encoder. After pretraining, we utilize the unified feature representations output by the encoder as the input for the downstream task.

mechanism called *"Shared Encoder"* to learn the general feature representations for multisource HSIs. Unlike other feature representation methods, *"Shared Encoder"* have some new features, the main contributions are as follows: 1) it can handle spectral data of arbitrary spectral dimensions, ranges, and resolutions; 2) it is able to project multisource HSIs into the same feature space through self-supervised learning; 3) it has the capability to fully leverage existing multisource HSIs; 4) it demonstrates strong generalization performance and can enhance performance in the classification task.

II. Methodology

The shared encoder aims to provide a unified representation of multisource HSIs, effectively adapting to spectral signals of varying dimensions. Consider 3-D HSI as $\mathbf{H} \in \mathbb{R}^{h\times w\times b}$, where h, w, and b represent the height, weight, and bands, respectively. Since spatial information is not considered in our method, we flatten the spatial dimensions: $\mathbf{H} \to \mathbf{X} \in \mathbb{R}^{n\times b}$, where $n = h\times w$. The input to the model is a spectral signature $x \in \mathbb{R}^{1\times b}$. The overall architecture of the shared encoder is illustrated in Fig. 1. It primarily comprises a bottleneck 1-D convolutional AE, consisting of a downsampling encoder and an upsampling decoder. The downsampling rate r is defined as $2^N (N = 1, 2, 3, \ldots)$, where the depth of downsampling module d is determined by r, specifically $d = \log_2 r$. To ensure that the signal can be evenly downsampled without remainders, we have designed an adaptive cropping layer as the input layer of the model. The cropping band number is computed as: $\Delta b = b - \lfloor b/r \rfloor \times r$. The resultant spectral signature can now be divided by r without any residual error. In the encoder section, the downsampling module f_D

$$f_D : \mathbb{R}^{1\times(b-\Delta b)} \to \mathbb{R}^{c\times\hat{b}} \tag{1}$$

and a global pooling layer f_{GP}

$$f_{GP} : \mathbb{R}^{c\times\hat{b}} \to \mathbb{R}^{c} \tag{2}$$

f_{GP} is used to extract the most significant factors and eliminate the dimension differences across multisource HSIs, where $\hat{b} = (b - \Delta b)/r$ and c is the output channel of f_D which can be preset. To extract higher-level features, we construct a bottleneck-structured MLP, the bottleneck feature is selected as the feature vector, which is utilized for subsequent tasks. On the decoder side, a repetitive padding layer f_R and an upsampling module f_U are used to perform the reverse operations corresponding to the encoder

$$f_R : \mathbb{R}^{c} \to \mathbb{R}^{c\times\hat{b}} \tag{3}$$

$$f_U : \mathbb{R}^{c\times\hat{b}} \to \mathbb{R}^{1\times(b-\Delta b)} \tag{4}$$

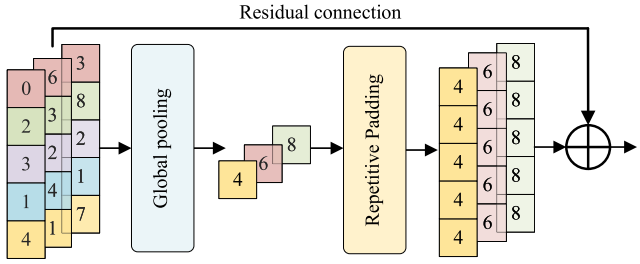

Fig. 2. Computational details between global pooling layer and repetitive padding layer.

f_R repeats the feature vectors after global pooling along the length dimension according to the saved parameter $\hat{b}$. f_U upsamples the feature vectors and parses the original HSI spectral signatures. Furthermore, we perform a residual connection between the features prior to pooling and the replicated features, aiming to supplement the detail deficiency caused by global pooling in the decoder, the computational details are illustrated in Fig. 2. Considering that the direction of the spectral vector in the feature space of HSIs is an important factor. In addition to utilizing MSE as the reconstruction loss, we incorporate the spectral angle mapper (SAM) as a model constraint which enables the model to learn information about the directionality of the spectral signal. SAM can be computed as follows equation:

$$S(x, y) = \cos^{-1}\left(\frac{xy^T}{||x||||y||}\right). \tag{5}$$

The shared encoder will be optimized as follows:

$$\theta^* = \arg\min_{\theta} \frac{1}{n}\sum_{i=0}^{n} |x_i^{\theta} - x_i|^2 + \alpha S\left(x_i^{\theta}, x_i\right) + \beta||\theta||_2^2 \tag{6}$$

where θ^* represents the optimal weights, x_i^{θ} represents the reconstructed spectra, $||\theta||_2^2$ represents the L2 regularization,

TABLE I
DETAILED INFORMATION OF THE DATASETS USED

No.	Dataset	Sensor	Band	Spectral Range
1	Botswana	Hyperion	145	400-2500 nm
2	Houston	ITRES CASI-1500	144	380-1050 nm
3	KSC	AVIRIS	176	400-2500 nm
4	NewXiongAn	PHI	250	400-1000 nm
5	Salinas	AVIRIS	224	400-2500 nm
6	Xuzhou	HYSPEX	436	415-2508 nm
7	WHU-Hi-HanChuan	Nano-Hyperspec	274	400-1000 nm
8	WHU-Hi-LongKou	Nano-Hyperspec	270	400-1000 nm
9	WashingtonDC	Hydice	191	400-2400 nm
10	Pavia University(PU)	ROSIS	103	430-860 nm
11	Indian Pines(IN)	AVIRIS	220	400-2500 nm
12	Chikusei	Hyperspec-VNIR-C	128	343–1018 nm

α and β represent the weight coefficients, by adjusting these coefficients can make the model's selection more focused on either the amplitude or the direction of the spectral signal. To demonstrate the model's robustness, the default value is set to 1.

III. EXPERIMENTAL RESULTS

A. Datasets Description

We utilized a total of 12 hyperspectral images from distinct sensors as experimental materials. The detailed information for these datasets is provided in Table I. Among them, datasets numbered 1–9 are employed as training sets to pretrain the shared encoder, while datasets numbered 10–12 are serve as the testing sets to evaluate the generality and the accuracy of the extracted features. For the classification task, to robustly showcase the cross-data source capability of the shared encoder, we select the unseen benchmark classification dataset PU and IN datasets as the training set. For IN, it is a challenging dataset owing to its low spatial resolution, thus we choose to increase the quantity of training data in this dataset, using 5% and 10% of the data as training sets. For PU, we use 1% and 5% of the data as training sets.

B. Experiment Settings

In our experiments, we primarily discuss two aspects: the accuracy and generalization ability of feature representation, as well as the enhancement of classification performance by pretrained model. PSNR and SSIM metrics are utilized to measure the similarity between reconstructed and original HSIs. Overall accuracy (OA), average accuracy (AA), and the Kappa coefficient are employed as performance metrics for classification tasks. Since our method currently only considers spectral information, we select SVM [16], neural network (NN) [17], 1D-CNN [18], and MSR [12] which only utilize spectral information for comparison. All experiments were conducted on the same environment: Ubuntu20.04 LTS, GPU: NVIDIA GeForce RTX 4090 24GB, RAM:256 GB. AdamW is utilized as an optimizer, the learning rate was set to 0.001 and the weight decay was set to 0.0005.

C. Discussion

Table II presents our reconstruction results on muti-source HSIs from different and unknown sensors. From the table, it can be observed that the model achieves the best performance on three testsets when $r = 32$, the PSNRs are 27.07, 22.27, 29.35 dB, and the SSIMs are 0.93, 0.82, and 0.88, respectively. Besides, the shared encoder also achieved satisfactory reconstruction accuracy at other downsampling rates. Figs. 3, 4, and 5 display the original images and their reconstructions for a subset of bands, along with their respective error maps. This demonstrates that our model can

TABLE II
EVALUATION OF THE RECONSTRUCTION ACCURACY OF SHARED ENCODER

Shared Encoder	PaviaU		Indian Pines		Chikuse	
	PSNR $[dB]$	SSIM []	PSNR $[dB]$	SSIM []	PSNR $[dB]$	SSIM []
$r = 8$	25.27	0.76	20.32	0.67	27.49	0.73
$r = 16$	25.87	0.81	21.64	0.74	28.33	0.79
$r = 32$	**27.07**	**0.93**	**22.27**	**0.82**	**29.35**	**0.88**

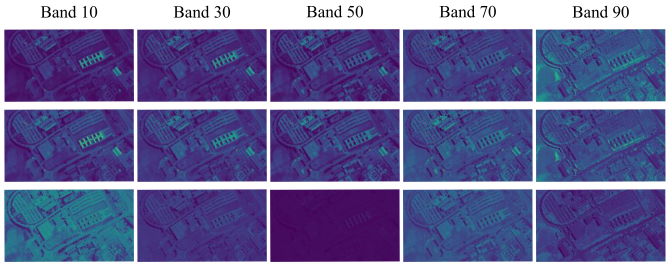

Fig. 3. Reconstruction results of PU. The first row is the original image, the second row is the reconstructed image, and the third row is the error map.

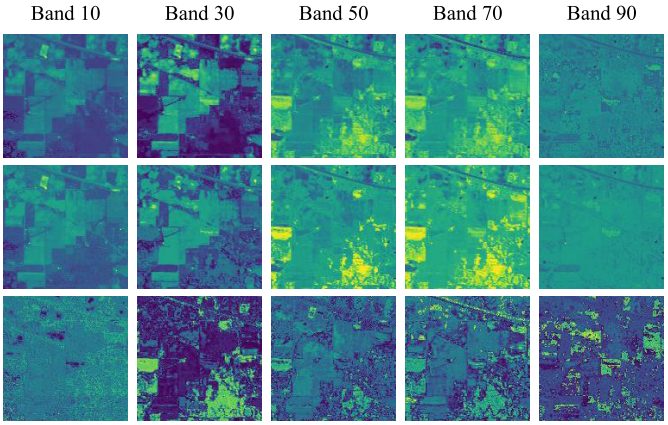

Fig. 4. Reconstruction results of IN. The first row is the original image, the second row is the reconstructed image, and the third row is the error map.

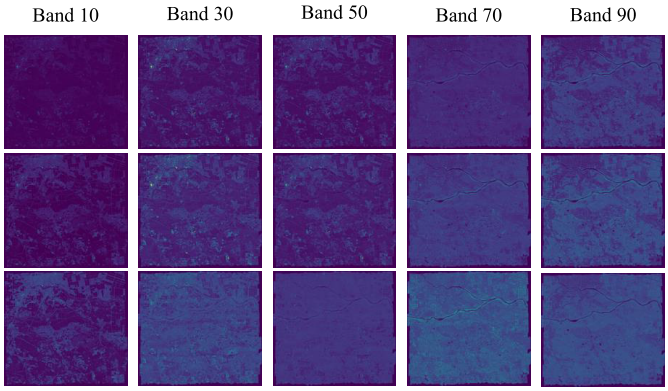

Fig. 5. Reconstruction results of Chusikei. The first row is the original image, the second row is the reconstructed image, and the third row is the error map.

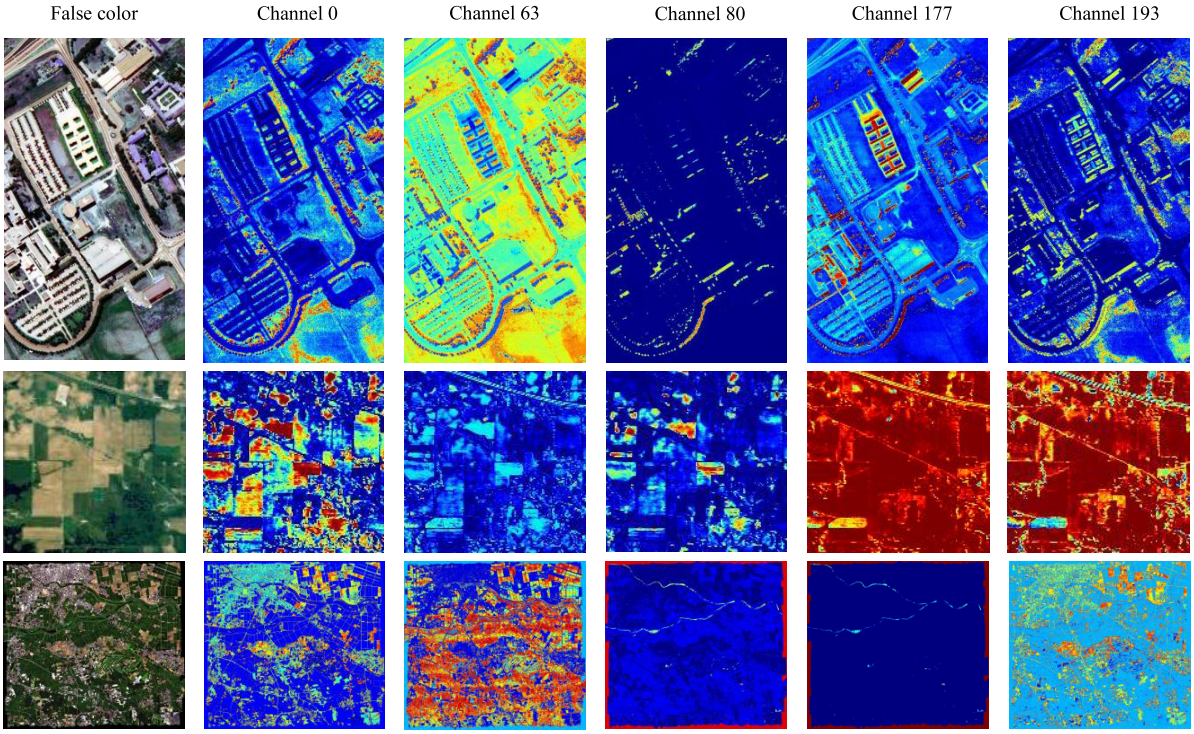

Fig. 6. Visualizations of the subset of learned representations. The first column shows the false color images of the three reconstruction test datasets, while the second through sixth columns provide visualizations of the feature representation channels 0, 63, 80, 177, and 193.

TABLE III
METRICS AND INFERENCE TIME OF CLASSIFICATION TASK

Methods	1% PU				5% PU				5% IN				10% IN			
	OA [%]	AA [%]	Kappa [%]	Time [s]	OA [%]	AA [%]	Kappa [%]	Time [s]	OA [%]	AA [%]	Kappa [%]	Time [s]	OA [%]	AA [%]	Kappa [%]	Time [s]
SVM [16]	75.86	60.49	65.72	3.81	78.57	64.14	69.88	14.97	36.93	13.11	20.29	**0.20**	47.84	24.52	36.06	0.80
NN [17]	83.70	78.38	78.26	**1.15**	91.17	86.83	88.25	**1.11**	76.60	**70.96**	73.11	0.38	82.78	80.66	80.35	**0.35**
1D-CNN [18]	83.76	78.73	78.14	1.31	93.37	89.93	91.15	1.21	76.52	70.88	73.06	0.42	84.94	80.89	82.79	0.39
MSR+ResNet [12]	79.16	71.64	71.87	50.00	86.95	82.21	82.49	46.95	49.42	26.81	39.68	29.10	58.97	39.77	52.37	27.92
* W/O Pretrian	86.11	77.81	81.53	1.31	92.19	87.53	89.58	1.24	70.43	61.86	66.15	0.52	79.97	68.67	77.14	0.50
* Pretrain	**87.51**	**81.12**	**83.23**	1.31	**93.54**	**90.39**	**91.43**	1.26	**78.71**	70.53	**75.66**	0.53	**85.24**	**82.00**	**83.15**	0.50

* indicates the shared encoder, W/O is the abbreviation of without. The optimal results are indicated in bold, and the second-best results are underscored.

extract highly reliable universal features and possesses good generality through a large amount of pretraining data. Error maps reveal that there are still some differences between the reconstructed and original data. This is because, similar to meta-learning, our goal is to project all types of HSIs into a unified representation space where each maintains its original information reasonably well rather than achieving the best performance on each task.

Furthermore, we conducted a visualization analysis of the learned representation vectors, with visualizations of randomly selected channels shown in Fig. 6. For PU, it is evident that channels 0, 63, and 193 exhibit higher activation values in vegetated areas, while channel 80 shows higher activation values in shadowed regions, and channel 177 exhibits higher activation values in built-up areas. For IN, channel 0 exhibits higher activation values in recently seeded land, channel 63 shows higher activation values in vegetated areas, channel 80 indicates higher activation values in areas yet to be cultivated or lightly cultivated, while channels 177 and 193 are less sensitive to wheat. For Chikusei, channels 0 and

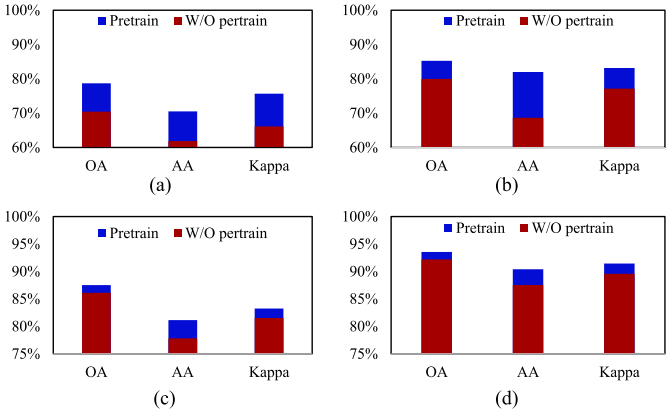

Fig. 7. Effects of pretrained model. (a) 1% IN (b) 5% IN (c) 5% PU (d) 10% PU.

193 exhibit higher activation values in recently seeded land, channel 63 shows higher activation values in vegetated areas, channel 80 exhibits a weak response to water bodies and areas with high moisture content, while channel 177 responds only

to water bodies. It can be observed that, despite the input data of the model originating from different sensors and scenes, there is a certain correlation among the information focused on by each dimension of the representation vectors, and different channels respond differently to different materials. This indicates that our model has learned the intrinsic relationships of multisource HSIs during pretraining and generates unified implicit features of HSIs.

Comparison and ablation results were conducted to assess the promoting effect of our method on downstream task performance, the results are presented in Table III. It shows that the pretrained model outperformed others across all experimental conditions, except for the AA metric of the 5% IN dataset, where it trails the best performance by only a marginal 0.43%. The inference time of the shared encoder is almost on par with 1D-CNN, but it significantly outperforms other methods in terms of performance, demonstrating the advancement of our method. The differences with and without pretraining are illustrated in Fig. 7. For IN, with 5% training samples, the OA, AA, and Kappa increased by 8.28%, 8.67%, and 9.51%, respectively; with 10% training samples, the metrics increased by 5.27%, 13.33%, and 6.01%, respectively. For PU, with a 1% training sample, the metrics showed increases of 1.4%, 3.31%, and 1.7%, respectively. Similarly, with 5% training samples, the metrics increased by 1.35%, 2.86%, and 1.85%, respectively. As comparison and ablation demonstrated, the pretrained model produces observable increases in all three metrics. This has been attributed to the fact that our pretraining model uses an extensive amount of multisource data, which allows it to sufficiently learn potential features and understand the internal logic of HSI, in contrast to existing strict models that only use small amounts of single-source data. Additionally, it naturally possesses cross-domain capabilities and is capable of transferring its learned information to new datasets with ease.

IV. Conclusion

In this article, we have designed a shared encoder that constructs a unified feature space for multisource HSIs through self-supervised learning. The model is highly compatible with HSIs from various sources, requiring no structural adjustments after pretraining. Experimental results show an acceptable precision in reconstruction tasks and a significant improvement on classification tasks across different datasets. It provides the possibility of large-scale pretrained models for HSI analysis. We are currently applying for access to data from multiple platforms. In future work, we aim to construct a fundamental model for multisource HSIs that takes into consideration spatial information as well as deterioration, noise effects, or variabilities experienced during the imaging process to improve the practical application of HSI.

References

[1] 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.

[2] 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.

[3] G. Taskin, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2918–2928, Jun. 2017.

[4] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[5] 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.

[6] J. M. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1829–1845, Mar. 2018.

[7] P. Duan, X. Kang, S. Li, P. Ghamisi, and J. A. Benediktsson, "Fusion of multiple edge-preserving operations for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10336–10349, Dec. 2019.

[8] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 7726–7738, Sep. 2020.

[9] P. Xiang, S. Ali, J. Zhang, S. Ki Jung, and H. Zhou, "Pixel-associated autoencoder for hyperspectral anomaly detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 129, May 2024, Art. no. 103816.

[10] S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, Jun. 2019.

[11] 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, 2024.

[12] X. Kang, Y. Zhu, P. Duan, and S. Li, "Two dimensional spectral representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2023, Art. no. 5502809.

[13] D. Hong et al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[14] J. Kuester, W. Gross, S. Schreiner, W. Middelmann, and M. Heizmann, "Adaptive two-stage multisensor convolutional autoencoder model for lossy compression of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5530022.

[15] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 3, 2024, doi: 10.1109/TPAMI.2024.3362475.

[16] Y. Chen, X. Zhao, and Z. Lin, "Optimizing subspace SVM ensemble for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1295–1305, Apr. 2014.

[17] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[18] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.