






S²MoINet: Spectral–Spatial Multiorder Interactions Network for Hyperspectral Image Classification

Yanan Jiang , Student Member, IEEE, Heng Zhou , Student Member, IEEE, Zitong Zhang , Student Member, IEEE, Chunlei Zhang , and Kai Zhang 

Abstract—Deep learning methods have shown great promise in automatically extracting features from hyperspectral images (HSIs) for classification purposes. Recently, researchers have recognized the importance of high-order feature interactions—capturing relationships between features in different image regions—in extracting discriminative features. Despite their effectiveness, the existing deep learning models for HSI classification often overlook high-order feature interactions, resulting in sub-optimal performance. To address this issue, we propose a novel spectral–spatial multiorder interaction network (S²MoINet) for HSI classification. The proposed framework can effectively extract highly discriminative features by leveraging correlations between features in different locations, significantly improving the classification accuracy. More specifically, we design a multiorder spectral–spatial interaction block in the framework to extract the high-order and generalized features by leveraging the interaction between spatial and spectral features. Based on experimental results from four public HSI datasets, it has been shown that the proposed S²MoINet delivers optimal classification results when compared to other state-of-the-art methods.

Index Terms—Hyperspectral image classification, multiorder interaction, neural networks, recursive gating mechanism, spectral–spatial feature representation.

I. INTRODUCTION

A HYPER SPECTRAL image (HSI) contains numerous spectral bands that cover a specific, continuous, and narrow wavelength range of the electromagnetic spectrum, rendering them abundant in spectral information [1]. Standard processing methods for the HSI include noise reduction, feature selection, and classification [2], [3], [4]. HSI classification is a fundamental task of HSI interpretation due to its ability to help researchers identify specific materials, and this technique has been used successfully in a variety of applications, such as urban planning [5], military reconnaissance [6], agricultural

production [7], earth observation [8], and mineral resource prospecting [9].

In the early stages of HSI classification research, numerous explorations have been conducted using machine-learning-based models [10]. These methods typically require manual feature extraction and utilize algorithms, such as support vector machines, random forests, and k -nearest neighbors, to implement HSI classification [11]. Regrettably, these methods meet difficulty in effectively capturing high-dimensional and nonlinear features in HSI data, leading to suboptimal classification accuracy.

The classification of HSI has been widely researched with deep learning models thanks to the development of computing technology and algorithms. Among these deep learning models, convolutional neural networks (CNNs) have been extensively explored for their automatic capability in extracting HSI features [1], [12]. In particular, various types of CNNs have been devised, including capsule networks [13], [14], graph CNNs [15], [16], and morphological convolution network models [17]. These models improve object feature representation, topological feature relationships, and morphological description by leveraging the flexible feature extraction of convolution. CNNs allow for extracting abundant spectral–spatial features from HSI data, making them a popular choice for HSI classification tasks. Nevertheless, the shared convolution operators of these models tend to create homogeneous features, which cannot effectively capture the discrepancy information of different spectral channels and spatial locations in the HSI. As a result, the capacity of models to discriminate HSI features is severely suppressed.

The attention mechanism is a feasible way to enhance feature distinguishability. It simulates human visual perception by selectively focusing on salient parts, rather than treating every part equally [18]. More studies have been conducted to introduce attentional mechanisms into the HSI classification task to obtain more representative features from the perspective of channel attention, spatial attention, and spectral–spatial attention [19], [20], [21]. Furthermore, methods such as multiscale attention block [22], [23] and adaptive attention block [24] have been designed to obtain more discriminative and comprehensive HSI spectral–spatial feature information.

The self-attention (SA) mechanism engages with features from various locations, leading to the development of several SA-mechanism-based approaches for HSI classification. SA mechanisms, grounded in this principle, determine feature

Manuscript received 4 May 2023; revised 25 June 2023; accepted 18 July 2023. Date of publication 25 July 2023; date of current version 7 August 2023. (Corresponding author: Heng Zhou.)

Yanan Jiang is with the School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China (e-mail: yananjia@mail.bnu.edu.cn).

Heng Zhou and Kai Zhang are with the College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China (e-mail: zhou_ac@163.com; 541295859@qq.com).

Zitong Zhang is with the School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China (e-mail: 3001200116@email.cugb.edu.cn).

Chunlei Zhang is with the Beijing Zhongdi Runde Petroleum Technology Company, Ltd., Beijing 100083, China (e-mail: zcl_3559@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3298477

weights based on their similarity to other features [25], [26]. Recent studies have demonstrated this through various attention-based methods, such as the spectral attention mechanism, the SA context network, the spectral–spatial self-mutual attention network, etc. [27], [28], [29], [30], [31]. The SA mechanism can only obtain the second-order interaction information of the data, and its operation requires the calculation of the similarity matrix between all the pixels. It may lead to high computational complexity when utilizing it for the HSI classification tasks, resulting in inefficient models.

Similarly, the gating mechanism works by assigning weights to features, which helps to suppress redundant information and highlight relevant information. This ultimately leads to improved performance in the target task [32], [33], [34]. More recently, Rao et al. [35] introduced HorNet for computer vision tasks, offering a novel perspective on feature extraction wherein feature interaction entails capturing relationships between features across distinct image regions. They categorized the feature extraction approaches of the CNN [12], SENet [36], and visual transformer [26] as zeroth-order, first-order, and second-order interaction features, respectively. Specifically, they employed recursive gating convolutions to eliminate excessive information, thereby preserving distinguishable features through iterative evaluation of correlations across various locations and assigning negligible weights to homogeneous features. The application of high-order feature interactions to enhance RGB image classification effectiveness has been demonstrated across multiple domains [37]. HorNet’s success in computer vision tasks has inspired us to develop HSI classification models. However, the existing deep learning approaches for the HSI have not incorporated multiorder interaction features, which is a limitation that hinders the performance of these models. Moreover, owing to the unique spectral characteristics of HSI data, directly applying RGB image classification methods to HSI classification presents a challenge.

To tackle the abovementioned challenges, we develop a novel classification framework for the HSI, named spectral–spatial multiorder interaction network. This framework has two main parts. First, we introduce a spectral–spatial feature representation and fusion (S^2FRF) block, which extracts spectral–spatial feature information from the input HSI patches. Second, we design a multiorder spectral–spatial interaction (MoS^2I) block for high-order feature extraction, which is composed of the multiorder spectral–spatial gating mechanism (MoS^2GM) module together with other common operations to acquire the local and global high-order contextual interaction feature. Through combining the high-order spectral–spatial interaction features, our approach can obtain a more sufficient high-order information representation of HSI objects and then can effectively identify more complex targets in HSI data. Meanwhile, considering the feature importance of spectral and spatial domains, we adopt the SA mechanism to emphasize the interested information, thus allowing the model to suppress redundant information effectively, especially in spatial–spectral features. Furthermore, the skip connection is also employed to transmit low-order information into the deep layer for capturing high-order interaction features. The main contributions of this article can be summarized as follows.

- 1) The main innovation of this article is to introduce the idea of high-order spectral–spatial feature extraction and interaction into the HSI classification task.
- 2) According to the properties of HSI data, a spectral–spatial multiorder interaction network ($S^2MoINet$) is designed to achieve effective and efficient HSI classification by extracting high-order and generalized features based on the interaction of spatial–spectral features.
- 3) To comprehensively represent spectral–spatial features and their multiorder interactions in the spectral and spatial domains, we introduce a novel MoS^2I block that employs a gating mechanism to iteratively suppress redundant features.
- 4) We have conducted both the qualitative and quantitative assessments of the classification capabilities of our $S^2MoINet$ on various HSI datasets, with thorough ablation studies. The experimental results indicate that our proposed network outperforms other current backbone networks with a maximum improvement of 10% in overall accuracy (OA).

The rest of this article is organized as follows. Section II summarizes the related work of HSI classification. Section III introduces the proposed $S^2MoINet$ model and describes its component modules in detail. Section IV presents the experimental settings and result analysis. Section V presents the discussion. Finally, Section VI concludes this article.

II. RELATED WORK

HSI data usually contain massive spectral bands and spatial pixel information, and the information interaction between spectral and spatial dimensions is essential for the analysis and application of hyperspectral data [38]. When performing HSI classification tasks, traditional models frequently use two primary kinds of spectral-based feature extraction methods and spectral–spatial-based feature extraction methods [39], [40]. In addition, the information in HSI data is also complex and diverse, and the shallow (texture) features of the HSI obtained through the model cannot express the information from ground objects well, so it is necessary to further capture the high-order (contour, shape, etc.) abstract features of the HSI to achieve adequate representation of the ground object information.

In this article, we introduce a novel model called the $S^2MoINet$, which integrates the spectral–spatial interaction relationship between features at different locations through an MoS^2GM . It can effectively capture the more representative and discriminative high-order interaction feature information of the HSI and finally achieve a more accurate classification of the HSI.

In this section, we briefly review the development of elementary ideas and specific operations associated with the proposed model, namely, the convolution-based model, attention mechanism, SA mechanism, and gating mechanism.

A. Convolution-Based Model

Deep learning techniques have recently gained attention for extracting spectral and spatial features simultaneously [1], [4], [41]. CNNs, a deep learning approach, automatically extract local structure and deep abstract features from input image

data. CNNs have been widely applied in computer vision tasks and have achieved remarkable performance in HSI spatial–spectral feature extraction and classification. For instance, Chen et al. [12] adopted the CNN approach to automatically obtain spectral–spatial features from the HSI, achieving better classification performance. Zhong et al. [42] proposed a spectral–spatial residual network to realize robust HSI classification results by utilizing the residual blocks and 3-D convolutions. Li et al. [43] utilized a fully group CNN method and achieved robust classification accuracy with relatively less parameters. Meng et al. [44] designed a residual dense asymmetric convolutional network that reduced feature redundancy and parameters through a novel concatenation mechanism and asymmetric convolution so as to capture discriminative HSI features.

In recent years, novel-convolution-based variants have emerged, including graph convolution networks [45], [46] that can deal with non-Euclidean structures, the morphological convolutional network [17] that can retain the basic features of images (such as the boundary, shape, and structure information), and the capsule network [14], [47] that can capture the pose and spatial relationship of objects. While CNN-based models have proven effective for HSI classification tasks, they have limited capacity for capturing long-range dependencies and interactions across different regions. Therefore, it is not enough to obtain spectral–spatial features by convolution operations alone.

B. Attention Mechanism

The attention mechanism enables the evaluation of the relative importance between different input features, obtains the correlation between data, and eventually obtains more discriminative features in the image [36].

Many experts have already utilized the attention mechanism to perform HSI classification tasks. Wang et al. [19] designed a model that utilized the squeeze-and-excitation module to acquire more conducive spectral and spatial features for HSI classification. Furthermore, experts also studied the attention mechanism and the feature fusion block, which can learn discriminative features containing HSI spectral information and spatial background features from different sensory fields [20], [24]. Zhu et al. [21] adopted a residual spectral–spatial attention model to suppress useless band and spatial information for adaptive feature selection and refinement of HSI spectral–spatial information. Gao et al. [22] proposed a densely connected multiscale attention model to effectively emphasize the features and improve the extraction and fusion capability of HSI spectral–spatial features. Fang et al. [23] constructed a network by utilizing the multiattention method to effectively fuse the spectral–spatial features for superior classification results.

Compared to the previous convolution-based models, the attention mechanism can effectively capture the representative features in HSI data, which is conducive to HSI classification.

C. SA Mechanism

The SA mechanism allows each element of an input sequence to interact with others by calculating the scaled dot-product of query, key, and value, which can facilitate the learning of

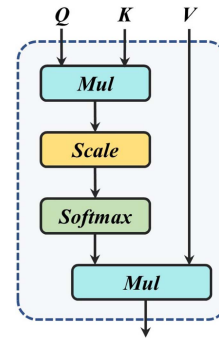


Fig. 1. Structure of the SA mechanism, where Q , K , and V are the input matrices. Mul stands for the matrix multiplication operation, $Scale$ represents the scaling operation, and $Softmax$ is the normalized exponential function.

correlations between different parts of the data [25]. The basic structure details are shown in Fig. 1. Let $X \in \mathbb{R}^{N \times D}$ denote the input data, and the SA layer is calculated as follows:

$$SA(Q, K, V) = Softmax(QK^T/\sqrt{d})V \quad (1)$$

where d is the hidden dimension of queries Q and keys K . The query $Q = XW^Q$, key $K = XW^K$, and value $V = XW^V$ matrices are computed via linear projections.

The correlation information between data can be obtained by matrix multiplication operation. Therefore, it can dynamically generate weights to effectively capture second-order interaction information in data after two successive matrix multiplication operations in SA.

For HSI classification, an increasing number of studies have proposed using SA mechanisms to enhance the ability of models to capture remote spectral–spatial interactions across different bands and spatial locations [48]. Specifically, Qing et al. [27] constructed an end-to-end transformer model by utilizing the SA mechanism to effectively capture the long-range continuous spectrum relationship in HSI. Zhou et al. [31] adopted the self-mutual attention mechanism and the SA module to extract spectral–spatial correlations and account for long-range dependencies. Zhang et al. [49] introduced an innovative attention model called the global-local block spatial–spectral fusion, which was designed to effectively gather information from both the spectral and spatial dimensions to classify HSI data. Recent studies have shown that models combining convolution and SA mechanisms [29], [50] generally outperform single-model approaches.

D. Gating Mechanism

The gating mechanism is also a wide application in deep learning for image feature extraction. It can perform the gating operation on the convolution or recurrent layer and selectively amplify or suppress feature channels, dimensions, or time-series features, which enhances the adaptability to different inputs. For example, Mou et al. [32] regarded the HSI as sequence data and then employed the recurrent neural network (RNN) model along with the custom activation function and adjusted gated recurrent unit to address the multiclass classification problem of the HSI.

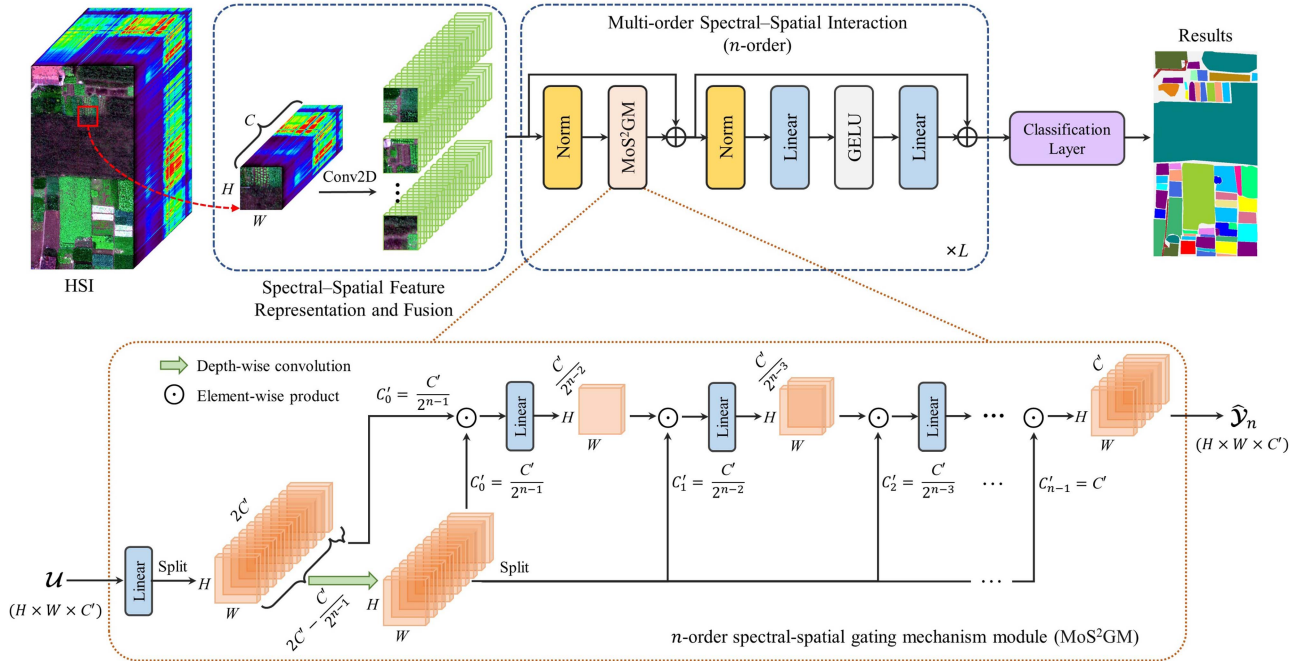


Fig. 2. Overall structure of the S^2 MoINet model for HSI classification, which includes the S^2 FRF block to capture the spectral–spatial information in the HSI, the multiorder spectral-space interaction block that can obtain arbitrary-order spectral–spatial interaction information, and the classification layer for ultimate classification purpose. Details of the n -order spectral–spatial gating mechanism module are shown in the bottom half of the figure.

Hang et al. [51] developed a cascaded RNN model with gated recurrent units to analyze features and consider complementary information of HSI data, which allowed the model to obtain more discriminative spectral–spatial information for enhanced classification performance. Zhou et al. [52] introduced a multiscanning strategy with the RNN, which incorporated a gating mechanism to capture the sequential feature of HSI pixels and extensively account for the spatial correlation in HSI patches. Prior research underscored the ability of recursive gating mechanisms to refine information flow, enhance the discriminability and generality of features, and ultimately achieve superior model classification performance.

More recently, the recursive gating mechanism has been introduced and demonstrated to effectively facilitate complex feature representation in various neural network models. For HSI classification tasks, incorporating the recursive gating mechanism allows for filtering and regularizing features to obtain high-order interaction information, thereby enhancing classification performance [53]. Meng et al. [54] presented a network for feature fusion that utilized spatial attention and gated mechanisms to highlight discriminative regions in multilayer feature maps and extracted key areas information with high efficiency.

It is crucial to consider the high-order information of the input data for the classification task. Therefore, Li et al. [37] applied the high-order feature interaction to improve optical remote sensing scene classification performance. Subsequently, Rao et al. [35] presented high-order feature interactions using the recursive gating mechanism. Regrettably, these methods focus solely on high-order features in RGB images, presenting challenges when introducing high-order features directly to HSI classification, particularly in efficiently handling the fusion of

spectral–spatial information. Therefore, we combine the SA mechanism to get the interactive information of the HSI and utilize the recursive gating mechanism to solve the problem of capturing high-order features at the root.

III. METHODOLOGY

In this section, we present the S^2 MoINet framework, which emphasizes spectral–spatial interactions between pixels and their neighboring regions, enabling arbitrary MoS²I to facilitate HSI classification. The overall classification structure is displayed in Fig. 2. It is comprised of three primary components.

- 1) The S^2 FRF block, which is performed for the HSI to obtain the output spectral–spatial features simultaneously: Meanwhile, it also considers the local contextual information and effectively preserves features regarding local spatial neighbors. Section III-A contains the necessary information regarding the spectral–spatial feature embedding.
- 2) The MoS²I block, which contains several MoS²GM modules: The MoS²GM module can realize the modeling of multiorder interaction between spectral and spatial domains and obtain the corresponding multiorder encoded feature output. Section III-B provides a comprehensive overview of this particular component.
- 3) The classification layer that utilizes the softmax function to accurately determine the probability of the input being categorized into a specific class, and it is described in Section III-C.

The original HSI input data can be regarded as 3-D data cube, which is specifically expressed as $\mathcal{X} \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ is the input spatial size and C is the spectral band

number. For the pixel located at any spatial position (i, j) in the input data \mathcal{X} , we can obtain the spectral vector $\mathbf{x}_{i,j} = (x_{i,j}^1, x_{i,j}^2, \dots, x_{i,j}^C)$, where $x_{i,j}^c \in \mathbb{R}$ ($i = 1, 2, \dots, H$, $j = 1, 2, \dots, W$, $c = 1, 2, \dots, C$) corresponds to the pixel at spatial position (i, j) in the band c of the HSI. Spectral vectors can carry a significant quantity of useful spectral fluctuation information that can be utilized to distinguish different ground objects. Moreover, each pixel and its neighboring pixels also contain abundant spatial information, such as the arrangement of ground object and the relationship with other objects in the HSI [31]. Therefore, it is very important to fully consider the spectral–spatial information simultaneously of HSI data in a certain neighborhood of each pixel when performing feature extraction.

A. Spectral–Spatial Feature Representation and Fusion

To fully obtain the spectral–spatial feature information from the HSI, we employ 2-D convolution to process the original input data of the HSI to ensure that the complex features and changes in the spectral and spatial domains are fully captured and merged, which we call the S²FRF block. Specifically, the convolution operation is performed on the input vector $\mathbf{x}_{(i,j)}$ of the original HSI at spatial position (i, j) to obtain its corresponding output features. The value $u_{(i,j)}^{k_2}$ of the k_2 th output channel at the spatial position (i, j) is calculated as follows:

$$u_{(i,j)}^{k_2} = \sum_{k_1=1}^C \sum_{\alpha=-m}^m \sum_{\beta=-m}^m w_{(\alpha,\beta)}^{k_1,k_2} x_{(i+\alpha,j+\beta)}^{k_1} + b_{k_2} \quad (2)$$

where $x_{(i+\alpha,j+\beta)}^{k_1}$ represents the value of input vector at the corresponding position in the k_1 th channel, C is the number of channels for input data, $2m+1 \in \mathbb{Z}^*$ is the size of convolutional kernels, the weight of the convolutional kernel at position (α, β) is represented by $w_{(\alpha,\beta)}^{k_1,k_2}$, and b_{k_2} is the corresponding bias. Finally, we can obtain a series of spectral–spatial feature embedded outputs $\mathcal{U} = \{u_{(i,j)}^{k_2} | i = 1, 2, \dots, H; j = 1, 2, \dots, W; k_2 = 1, 2, \dots, C'\} \in \mathbb{R}^{H \times W \times C'}$ from HSI data.

B. Multiorder Spectral–Spatial Interaction

To further acquire the intrinsic multiorder information in the HSI, we perform feature extraction modeling of the contextual information and MoS²I features on the outputs \mathcal{U} , which is obtained from the S²FRF block. Unlike the existing high-order feature extraction models, we consider the fusion of spatial–spectral interaction features for the properties of HSI data and also effectively introduce the MoS²GM module with locality perception and context aggregation capability and some other simple modules. Specifically, the linear projections, depthwise convolutions, and elementwise products are organically integrated to achieve a similar function of input-adaptive spectral–spatial mixing to SA. Meanwhile, the recursive gating operation is also performed on the above features to effectively capture the semantic interaction information between contextual data of the HSI. Through the above operations, we can construct an MoS²GM module to successfully obtain the contextual and

MoS²I information of the HSI. The MoS²GM module’s specifics are illustrated in Fig. 3.

For the first-order spectral–spatial gating mechanism, we utilize a input linear projection operation to perform channel mixing for the input feature token $\mathcal{U} \in \mathbb{R}^{H \times W \times C'}$; it can be written as follows:

$$\left[\mathcal{V}_0^{H \times W \times C'}, \mathcal{S}_0^{H \times W \times C'} \right] = \Psi_{\text{in}}(\mathcal{U}) \in \mathbb{R}^{H \times W \times 2C'} \quad (3)$$

where $\Psi_{\text{in}}(\cdot)$ is the linear projection function to perform channel mixing. \mathcal{V}_0 and \mathcal{S}_0 are the results obtained by linear projection. Note that, here, we split the result into \mathcal{V}_0 and \mathcal{S}_0 in order to perform the interaction of adjacent features.

Then, we conduct another depthwise convolutional operation $f(\cdot)$ on the output \mathcal{V}_0 and \mathcal{S}_0 to obtain the features \mathcal{V}_1 that considering the first-order spectral–spatial interaction of HSI data; the first-order interaction equation is expressed as follows:

$$\mathcal{V}_1 = f(\mathcal{S}_0) \odot \mathcal{V}_0 \in \mathbb{R}^{H \times W \times C'} \quad (4)$$

where the operator \odot denotes the elementwise product operation.

Finally, one more output linear projection operation $\Psi_{\text{out}}(\cdot)$ to perform channel mixing is utilized to make the first-order spectral–spatial interaction feature more robust and reduce the overfitting problem. Thus, the output feature vector $\hat{\mathcal{Y}}_1$ can be obtained by using the first-order spectral–spatial gating mechanism, as follows:

$$\hat{\mathcal{Y}}_1 = \Psi_{\text{out}}(\mathcal{V}_1) \in \mathbb{R}^{H \times W \times C'} \quad (5)$$

The similar feature extraction operation is also performed in the n -order ($n \in \mathbb{Z}^+$) spectral–spatial gating mechanism. First, the linear projection operation is performed on the input features obtained by the S²FRF block to acquire a set of projection features \mathcal{V}_0 and $\{\mathcal{S}_t\}_{t=0}^{n-1}$

$$\left[\mathcal{V}_0^{H \times W \times C'_0}, \mathcal{S}_0^{H \times W \times C'_0}, \dots, \mathcal{S}_{n-1}^{H \times W \times C'_{n-1}} \right] \\ = \Psi_{\text{in}}(\mathcal{U}) \in \mathbb{R}^{H \times W \times 2C'} \quad (6)$$

where $C'_t = C'/2^{n-t-1}$, $t = 0, 1, \dots, n-1$, represents the t th order of the MoS²GM module. The adjacent features are split into \mathcal{V}_0 and $\{\mathcal{S}_t\}_{t=0}^{n-1}$ to further capture the MoS²I features between them. Moreover, this splitting strategy utilized in the MoS²GM module enables the acquisition of features from coarse to fine, thus improving the capability of multiorder feature acquisition while reducing the computational complexity to some extent.

Since the arbitrary order interaction of spectral–spatial information for the HSI data can be captured by means of concatenating feature vector $\{\mathcal{S}_t\}_{t=0}^{n-1}$ and performing depthwise convolution, we recursively perform the depthwise convolution $f_t(\cdot)$ and dimension-matching operation $g_t(\cdot)$ of different orders for n times, so as to get the n th-order spectral–spatial interaction features. To stabilize the training process, we also scale the obtained output for each recursion by $1/\delta$ for obtaining stable output values

$$\mathcal{V}_{t+1} = f_t(\mathcal{S}_t) \odot g_t(\mathcal{V}_t) / \delta \in \mathbb{R}^{H \times W \times C'_t} \quad (7)$$

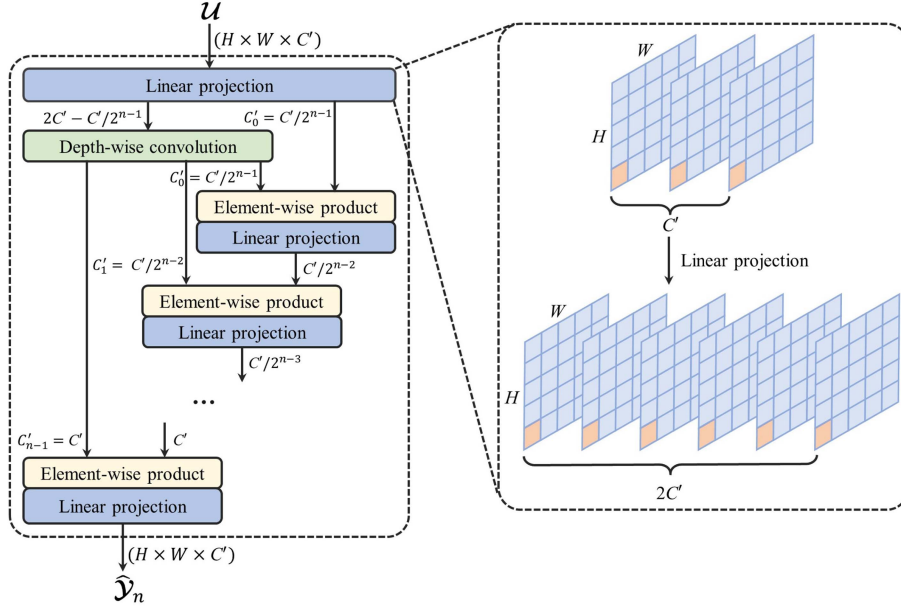


Fig. 3. Diagram of the n -order spectral-spatial gating mechanism module. The module utilizes linear projection to mix information in the spectral channel, depthwise convolution for feature extraction of HSI data in the spatial domain, and the elementwise product to effectively interact with spectral-spatial features.

where δ is a small real number to make the model training more stable. The dimension-matching operation $g_t(\cdot)$ can be calculated as follows:

$$g_t = \begin{cases} \text{Identity}, & t = 0, \\ \text{Linear}, & t = 1, 2, \dots, n-1 \end{cases} \quad (8)$$

where *Identity* represents the identity projection operation executed on the input channel C'_t , and *Linear* means to perform a linear projection operation on the input channel to make the C'_{t-1} and C'_t dimensions matching. We also utilize the final linear projection operation to obtain the output $\hat{\mathbf{Y}}_n$ of the n th-order MoS²GM by the following equation:

$$\hat{\mathbf{Y}}_n = \Psi_{\text{out}}(\mathbf{V}_n) \in \mathbb{R}^{H \times W \times C'} \quad (9)$$

where Ψ_{out} is the output linear projection to perform channel mixing, and \mathbf{V}_n is the result obtained through the last recursion operation of (7). We consider the output $\hat{\mathbf{Y}}_n$ in (9) as the result of n th-order spectral-spatial interaction.

The interaction between the features of HSI ground objects can better exploit the information associations present in HSI data, which, in turn, can improve the accuracy and robustness of HSI classification tasks. From the perspective of complex spectral-spatial feature interaction, in order to achieve a clearer expression, the calculation of the n th-order MoS²GM module can also be written as follows:

$$\mathcal{V}_{n(i,j)}^k = \sum_{(\alpha,\beta) \in \Theta_{i,j}} \sum_{c=1}^{C'} \mathcal{H}_{(i,j) \rightarrow (\alpha,\beta)}^c w_{\Psi_{\text{in}},(\alpha,\beta)}^{c,k} u_{(\alpha,\beta)}^c \quad (10)$$

$$\mathcal{H}_{(i,j) \rightarrow (\alpha,\beta)}^c = \mathcal{W}_{n-1,(i,j) \rightarrow (\alpha,\beta)}^k \mathcal{G}_{n-1,(i,j)}^k$$

where $\mathcal{V}_{n(i,j)}^k$ refers to the value of n th-order output feature (i.e., \mathbf{V}_n) at the spatial position (i, j) and the k th channel.

$\Theta_{i,j}$ is the local region centered at position (i, j) in the k th channel, and (α, β) is a point in this region. The arrow “ \rightarrow ” indicates the direction of connected mapping from the spatial position (i, j) to (α, β) . $\mathcal{H}_{(i,j) \rightarrow (\alpha,\beta)}^c$ is the result calculated from \mathbf{V}_{n-1} , including $(n-1)$ th-order interactions, $\mathcal{W}_{n-1,(i,j) \rightarrow (\alpha,\beta)}^k$ is the corresponding convolutional weight for $f_{n-1}(\cdot)$, and $\mathcal{G}_{n-1} = g_{n-1}(\mathbf{V}_{n-1})$ is the projection result of \mathbf{V}_{n-1} by (8). $w_{(\alpha,\beta)}^{c,k}$ represents the linear weight through $\Psi_{\text{in}}(\cdot)$ at the corresponding position, and $u_{(\alpha,\beta)}^c$ corresponds to the value at position (α, β) in the c th channel of the input feature token \mathbf{U} . Therefore, it can be seen from (10) that the above n th-order MoS²GM operation is equivalent to a general feature extraction and interaction module, which can perform better in considering the high-order spectral-spatial interaction features of the HSI.

When applying the MoS²GM module to HSI data, if the order is $n = 0$, the S²MoGM encoder module reduces to a standard 2-D convolution, capturing only simple spectral-spatial features and disregarding interactions within the HSI data. However, when the order is $n = 1$, the HSI data undergo the S²FE operation to obtain spectral-spatial features, followed by a matrix multiplication operation that determines the dynamic weight of input, accounting for the first-order spectral-spatial interaction of the HSI. When the order is $n = 2$, the MoS²GM module performs two consecutive matrix multiplications, equivalent to the SA mechanism, enabling extraction of spatial-spectral correlation information within the HSI data, thus considering the second-order spatial-spectral interaction. As the order n increases, the model accounts for MoS²Is in HSI data, effectively suppressing redundant information, particularly in spatial-spectral features, and facilitating subsequent HSI classification tasks.

The MoS²I output feature token $\hat{\mathbf{Y}}_n$ with equal sizes of input feature is obtained through the MoS²GM module. Then, after performing the normalizing operation, two subsequent linear projection operations and GELU activation function on the obtained token $\hat{\mathbf{Y}}_n$ to obtain the final output $\mathbf{Y}_{\text{MoS}^2\text{I}}$ of the MoS²I block. To get more detailed and multiorder features of the HSI ground objects, we can stack the MoS²I block for l times, where $l = 1, 2, \dots, L$.

C. Classification Layer

Next, the multiorder output features learned in the MoS²I block are fed into the classification layer. Specifically, the MoS²I feature tokens are initially pooled through the global average pooling operation. This is followed by a *Linear* projection operation, which generates the classification vector $\mathbf{q} = [q_1, q_2, \dots, q_B]^T$, and $q_b, b = 1, 2, \dots, B$, denotes the b th class of land cover. The calculation formula is as follows:

$$\mathbf{q} = \text{Linear}(\text{Pooling}(\mathbf{Y}_{\text{MoS}^2\text{I}})). \quad (11)$$

Then, the output vector $\mathbf{p} = [p_1, p_2, \dots, p_B]^T$ is calculated by the softmax function, so we can get the label of input data belonging to each type of ground object

$$p_b = \text{softmax}(q_b) = \frac{\exp(q_b)}{\sum_{i=1}^B \exp(q_i)}. \quad (12)$$

Let $\mathbf{z} = [z_1, z_2, \dots, z_B]^T$ represent the one-hot coding vector of the ground truth, where $z_b \in \{0, 1\}$ ($b = 1, 2, \dots, B$) denotes the b th class of the HSI; then, the proposed model's optimization procedure with the cross-entropy function $\mathcal{E}_{\text{loss}}$ can be calculated as follows:

$$\mathcal{E}_{\text{loss}} = - \sum_{b=1}^B z_b \log(p_b). \quad (13)$$

IV. EXPERIMENT AND ANALYSIS

A. HSI Dataset Description

We performed experiments on four HSI datasets, which are commonly used to evaluate the performance of HSI classification tasks. Fig. 4 provides false-color and ground truth maps information, while Table I presents a more specified description of the datasets.

- 1) *Indian Pines (IP)*: The first hyperspectral dataset was gathered by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) instrument in 1992 at the Indian Pines Northwestern Indiana [55]. The spatial resolution of IP is approximately 20 meters per pixel (mpp), and it has 145×145 pixels and 224 bands encompassing wavelengths from 400 to 2500 nm. During the experiment, bands that were absorbent were eliminated, and 200 bands were kept. The captured area mainly consists of various crops, irregular forests, and pastures, with a total of 16 different land cover types.
- 2) *Salinas Valley (SV)*: The second hyperspectral dataset was also gathered by the AVIRIS instrument in 1998 over Salinas Valley, CA, USA [56]. The spatial resolution

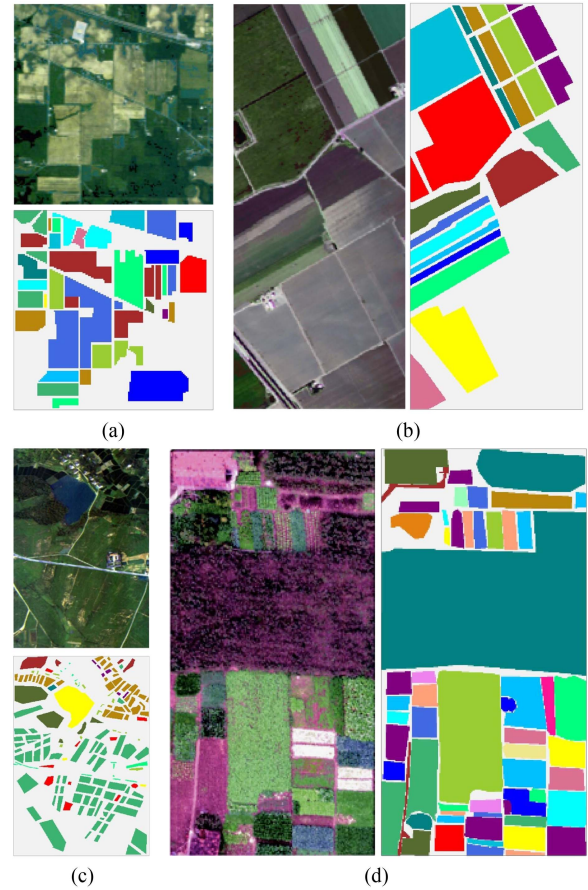


Fig. 4. False-color and ground truth maps of each HSI dataset. (a) IP. (b) SV. (c) TF. (d) HH.

of SV is approximately 3.7 mpp, and it has 512×217 pixels and 224 bands. Similarly, the absorbent bands were eliminated, with only 204 remaining. The capture area is dominated by regular fields with different crops, including 16 different land cover classes.

- 3) *Tea Farm (TF)*: The third hyperspectral dataset was collected by the Pushbroom Hyperspectral Imager instrument in 1999 over tea planting base in Fanglu Village, Changzhou City, Jiangsu Province, China [57]. The spatial resolution of TF is approximately 2.25 mpp, and it has 512×348 pixels and 80 bands covering wavelengths from 417 to 855 nm. The captured area includes ten different land-cover classes.
- 4) *HongHu (HH)*: The fourth hyperspectral dataset was collected by the unmanned-aerial-vehicle-borne instrument in 2017 over farming areas with various crop types in Hubei province, China [58]. The spatial resolution of HH is approximately 0.043 mpp, and it has 940×475 pixels and 270 bands covering wavelengths from 400 to 1000 nm. The captured areas include many types of crops in a complex agricultural scenes but also different varieties of the same crop, comprising a total of 22 different land-cover classes.

TABLE I
LAND-COVER CLASSES FOR THE DATASETS AND THEIR RESPECTIVE SAMPLE NUMBERS

No.	Color	Indian Pines		Salinas		Tea Farm		HongHu	
		Classes	Samples	Classes	Samples	Classes	Samples	Classes	Samples
C0		BackGround	10776	Background	56975	Background	124442	Background	59807
C1		Alfalfa	46	Brocoli GW1	2009	Masson P	5806	Reed roof	14041
C2		Corn N	1428	Brocoli GW2	3726	Bamboo F	2318	Road	3512
C3		Corn M	830	Fallow	1976	Tea P	28428	Bare soil	21821
C4		Corn	237	Fallow RP	1394	Reed	214	Cotton	163285
C5		Grass P	483	Fallow S	2678	Rice P	6809	Cotton F	6218
C6		Grass T	730	Stubble	3959	Sweet P	817	Rape	44557
C7		Grass PM	28	Celery	3579	Caraway	429	Chinese C	24103
C8		Hay W	478	Grapes U	11271	Weed	1861	Pakchoi	4054
C9		Oats	20	Soil VD	6203	Water body	6141	Cabbage	10819
C10		Soybeans N	972	Corn SGW	3278	Building/Road	911	Tuber M	12394
C11		Soybeans M	2455	Lettuce R4wk	1068			Brassica P	11015
C12		Soybeans C	593	Lettuce R5wk	1927			Brassica C	8954
C13		Wheat	205	Lettuce R6wk	916			Small BC	22507
C14		Woods	1265	Lettuce R7wk	1070			Lactuca S	7356
C15		Buildings GT	386	Vinyard U	7268			Celtuce	1002
C16		Stone ST	93	Vinyard VT	1807			Film CL	7262
C17								Romaine L	3010
C18								Carrot	3217
C19								White R	8712
C20								Garlic S	3486
C21								Broad B	1328
C22								Tree	4040
Total			21025		111104		178176		446500

TABLE II
DETAILS OF THE PROPOSED S²MOINET

No.	Block/Layer	Module	Operations	Hyperparameters
1	S ² FRF		Convolution	Kernel size = 3×3, Stride = 1, Input channel = C , Output channel = 64
		MoS ² GM	Linear projections depthwise convolution	Number of orders $n = 3$, Input channel = $C' = 64$, Output channel = 64
2	MoS ² I (×2)	Linear	Linear projection	Input channel = 64, Output channel = 128
		Linear	Linear projection	Input channel = 128, Output channel = 64
3	Classifier		Linear projection	Input nodes = 64, Output nodes = B

* C and B are the numbers of input channels and land-cover classes, respectively.

B. Experimental Settings

- 1) *Evaluation metrics*: Four quantitative evaluation metrics were introduced to quantitatively analyze the effectiveness of the S²MoINet and other models for comparison, including OA, average accuracy (AA), Cohen's kappa coefficient (κ), and the number of model parameters (Params).
- 2) *Comparison with state-of-the-art models*: For further analysis, we compare the S²MoINet method with seven commonly used and relevant deep architectures, including CNN [12], GhostNet [59], MAGCaps [60], ViT [26], Swin [61], SpectralFormer [29], ConvViT [62], and gMLP [63].
- 3) *Implementation details*: Validation experiments of the proposed model were conducted on a computer equipped

with 128-GB RAM and an NVIDIA RTX 3090 graphics card (24-GB VRAM) in the PyTorch 1.10 and Python 3.9 environment. We employed the cross-entropy loss function in the experiments. The AdamW optimizer was chosen, with a minibatch size of 64 and 300 training epochs. After every one-tenth of the total number of epochs (i.e., after epochs 30, 60, 90, etc.), the learning rate was multiplied by a factor of 0.9 to decrement from its initial value of 3e-4. Considering the different datasets, we utilized a third-order model with two times MoS²I block operation as an example. More details of the proposed model with base parameter settings are listed in Table II. In our studies, samples for the training and testing groups were drawn at random from the ground truth of the HSI dataset. We also selected samples with training ratios

TABLE III
CLASSIFICATION RESULTS FOR THE IP DATASET OF 15 × 15 INPUT SCALE AT 1% TRAINING RATIO

No.	CNN	GhostNet	MAGCaps	ViT	Swin	SpectralFormer	ConvViT	gMLP	S ² MoINet
C1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
C2	79.62	46.43	78.99	77.10	76.89	82.84	59.31	80.67	83.21
C3	72.29	47.11	59.88	51.81	57.35	75.54	65.66	56.75	69.95
C4	90.30	99.16	90.30	81.01	89.03	95.78	95.78	89.87	96.31
C5	69.57	76.60	85.51	66.67	71.01	71.43	78.88	73.91	66.70
C6	89.73	87.40	96.58	92.88	83.70	73.29	86.71	91.37	90.76
C7	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
C8	82.64	100.00	98.95	94.35	87.45	94.56	100.00	91.42	97.07
C9	100.00	100.00	100.00	100.00	100.00	95.00	100.00	100.00	100.00
C10	79.73	80.04	78.09	76.44	67.39	72.33	73.15	77.57	73.64
C11	95.19	82.12	77.76	87.70	85.30	91.53	76.74	92.95	93.34
C12	77.91	54.13	57.17	79.60	67.45	82.29	58.35	78.41	81.20
C13	100.00	100.00	100.00	97.07	100.00	100.00	100.00	100.00	100.00
C14	98.81	99.68	94.62	97.23	93.83	96.52	97.00	99.60	96.02
C15	90.16	88.60	80.31	78.50	94.30	86.27	81.87	91.45	96.13
C16	96.77	100.00	100.00	100.00	100.00	100.00	100.00	100.00	88.12
OA (%)	86.06	77.61	82.37	82.05	79.99	84.77	78.67	85.27	87.79
AA (%)	88.92	85.08	87.39	86.27	85.86	88.59	85.84	89.00	89.53
$\kappa \times 100$	90.93	73.84	78.89	87.99	86.60	85.39	75.16	90.29	91.37

The column that is highlighted in bold corresponds to the model that exhibits the best performance.

TABLE IV
CLASSIFICATION RESULTS FOR THE SV DATASET OF 15 × 15 INPUT SCALE AT 1% TRAINING RATIO

No.	CNN	GhostNet	MAGCaps	ViT	Swin	SpectralFormer	ConvViT	gMLP	S ² MoINet
C1	99.90	100.00	100.00	99.65	99.90	100.00	99.85	99.95	99.98
C2	100.00	99.87	100.00	100.00	99.92	98.66	99.76	100.00	100.00
C3	99.49	96.10	98.48	98.63	99.95	99.24	99.19	100.00	99.95
C4	99.14	99.57	97.49	94.12	95.05	94.98	96.77	98.71	99.17
C5	98.28	97.54	96.64	98.92	99.14	95.26	99.63	99.40	99.58
C6	100.00	100.00	99.85	100.00	100.00	99.57	100.00	100.00	100.00
C7	99.94	99.97	99.89	99.69	99.86	99.19	99.69	99.41	99.92
C8	93.24	97.80	94.89	94.64	96.51	92.26	94.08	95.05	95.76
C9	100.00	99.84	100.00	100.00	100.00	99.94	100.00	100.00	100.00
C10	98.08	99.91	96.98	98.93	98.20	98.32	99.15	98.60	98.54
C11	100.00	98.13	99.44	100.00	98.41	98.60	98.50	99.63	98.84
C12	99.79	99.79	99.79	99.53	100.00	100.00	98.34	100.00	99.90
C13	100.00	100.00	100.00	95.41	98.03	96.07	99.56	96.51	99.87
C14	99.91	96.26	97.20	98.88	99.91	97.57	99.81	99.91	99.79
C15	99.19	95.18	90.70	95.64	97.28	93.59	92.45	94.58	96.24
C16	99.94	99.94	99.28	100.00	100.00	96.13	98.67	99.89	99.88
OA (%)	98.83	98.59	97.28	98.25	98.76	97.33	97.62	98.66	98.85
AA (%)	99.18	98.74	98.16	98.38	98.88	97.46	98.47	98.85	99.21
$\kappa \times 100$	98.79	98.29	96.77	98.52	99.00	96.58	97.09	98.62	99.04

The column that is highlighted in bold corresponds to the model that exhibits the best performance.

of 0.5%, 1%, 1.5%, 2%, 2.5%, and 3% at random for each land-cover class in every dataset, and the remaining samples served as the testing data. We carefully chose to use half of the total number of samples as the training set for classes with small sample sizes. All the experiments were carried out ten times to ensure a fair comparison, and the average results are presented.

C. Classification Results in Different Models

For each dataset, Tables III–VI present the quantitative classification results in terms of the three metrics, i.e., OA, AA, and κ of each model on every HSI dataset. Among them, the labeled training sample is fixed at 1%, the scale of HSI input patches is selected as 15×15, and the classification results for each class and global metrics in the table are arranged by row, while the results for different models are displayed in columns.

It can be seen from tables that the S²MoINet model maintains a classification accuracy of 90% or above in each class compared with other models in most cases. By utilizing the S²MoINet model to extract MoS²I features, better access to deep abstract discriminative features in the data can be achieved. This is especially helpful as the resolution and structural complexity of HSI data increase, thus allowing the model to better classify the resulting complex HSI data. Therefore, the S²MoINet model acquires better classification results for each ground object class in multiple trials on different HSI datasets. In general, the proposed S²MoINet demonstrates the best results in terms of OA, AA, and κ , with all these metrics displaying the highest values. As compared to other models, the performance of OA, AA, and κ for the S²MoINet model has improved across all the datasets.

A qualitative evaluation was carried out, in which the classification maps generated by different models were visualized. Figs. 5–8 exhibit the obtained classification results for IP, SV, TF,

TABLE V
CLASSIFICATION RESULTS FOR THE TF DATASET OF 15×15 INPUT SCALE AT 1% TRAINING RATIO

No.	CNN	GhostNet	MAGCaps	ViT	Swin	SpectralFormer	ConvViT	gMLP	S ² MoINet
C1	100.00	99.52	99.35	99.59	99.53	99.04	97.40	99.97	99.60
C2	99.96	97.67	94.87	96.98	98.71	94.56	94.35	96.46	96.46
C3	99.74	99.65	98.92	99.35	99.63	99.37	98.62	99.66	99.70
C4	100.00	100.00	100.00	100.00	100.00	98.60	99.07	100.00	98.85
C5	100.00	99.72	99.84	99.75	99.49	99.90	98.93	100.00	99.84
C6	91.68	83.35	93.27	93.39	73.44	97.55	71.48	89.60	90.91
C7	99.53	98.37	98.60	100.00	99.53	95.80	100.00	98.83	99.91
C8	85.55	78.99	79.37	86.57	91.13	89.15	73.13	85.81	92.63
C9	99.67	98.70	99.77	100.00	99.98	99.98	99.30	100.00	99.95
C10	100.00	93.74	97.15	99.34	95.06	99.67	97.80	99.12	99.58
OA (%)	99.22	98.40	98.21	98.94	98.86	98.85	97.17	99.04	99.27
AA (%)	97.61	94.97	96.11	97.50	95.65	97.36	93.01	96.95	97.74
$\kappa \times 100$	99.50	97.59	97.36	99.30	99.27	98.83	95.74	99.37	99.58

The column that is highlighted in bold corresponds to the model that exhibits the best performance.

TABLE VI
CLASSIFICATION RESULTS FOR THE HH DATASET OF 15×15 INPUT SCALE AT 1% TRAINING RATIO

No.	CNN	GhostNet	MAGCaps	ViT	Swin	SpectralFormer	ConvViT	gMLP	S ² MoINet
C1	97.49	97.42	96.33	97.42	97.12	98.16	97.12	99.10	98.62
C2	97.75	95.19	84.74	91.69	84.77	95.76	88.38	95.42	95.63
C3	97.15	97.68	97.89	97.98	98.18	97.81	95.68	98.95	98.78
C4	99.91	99.88	99.92	99.88	99.75	99.91	99.85	99.62	99.84
C5	98.17	97.86	92.54	96.49	97.14	98.73	96.88	95.16	96.65
C6	99.48	99.25	99.10	99.47	99.52	99.34	99.34	99.62	99.37
C7	98.34	97.20	93.69	97.17	96.93	98.51	97.17	97.63	98.69
C8	89.57	89.05	57.13	90.73	89.76	96.13	86.98	86.24	93.05
C9	99.93	97.98	98.12	99.75	98.64	98.96	99.65	99.98	99.75
C10	97.88	97.99	94.46	97.93	98.60	97.51	96.39	97.61	98.95
C11	98.28	95.51	88.38	96.61	97.94	95.59	96.21	97.95	98.50
C12	96.26	93.34	83.86	95.91	96.79	94.90	91.38	97.00	97.08
C13	98.58	96.55	90.28	97.19	97.20	95.75	95.62	98.74	98.55
C14	99.40	93.22	93.12	97.38	97.62	98.18	93.23	98.45	98.36
C15	95.71	84.93	84.43	95.41	95.21	97.41	81.94	95.31	95.73
C16	98.61	98.09	97.12	97.73	99.37	99.23	97.62	98.69	99.51
C17	97.48	91.03	93.89	94.42	94.42	99.30	90.63	100.00	98.96
C18	95.83	96.27	91.14	91.73	95.46	96.30	93.78	93.47	96.93
C19	95.44	94.07	92.10	95.42	97.30	97.69	92.02	97.33	97.46
C20	95.64	95.41	91.88	92.34	93.29	98.11	91.02	96.56	96.61
C21	92.39	71.91	45.26	93.15	97.29	95.18	78.54	94.73	96.12
C22	97.52	98.74	88.69	97.55	98.89	99.41	98.42	99.65	98.89
OA (%)	97.42	97.92	96.01	96.56	96.86	97.97	97.56	97.52	98.04
AA (%)	97.13	94.48	88.82	96.06	96.42	97.63	93.54	97.14	97.82
$\kappa \times 100$	98.74	97.57	95.07	98.33	98.40	98.72	97.04	98.72	98.81

The column that is highlighted in bold corresponds to the model that exhibits the best performance.

and HH datasets at an input scale of 15×15 , with a training ratio of 1% and using various models. The results effectively show that, for most classes on each dataset, the proposed S²MoINet model surpasses the compared models. It employs a combination of SA and recursive gating mechanisms to construct the MoS²GM module, which effectively captures MoS²I features in HSI data. This powerful combination enables the model to effectively capture MoS²I features in HSI data, resulting in superior recognition performance for complex objects. Moreover, the gating mechanism plays a crucial role in mitigating the effects of noisy data points, leading to better accuracy in the HSI classification process. This also significantly reduces the number of misclassified sample points, effectively addressing the “salt and pepper phenomenon” that often arises during the HSI classification process. Overall, the proposed model delivers excellent classification results for each experimental dataset, demonstrating its generalization ability for HSI classification tasks.

The compared models display a considerable number of misclassified regions. Typically, the CNN-based models, such as CNN, GhostNet, and MAGCaps, tend to generate relatively smoother classification maps, thanks to their strong ability to nonlinearly fit data. As a recently used network architecture, the SA-mechanism-based models, such as ViT, Swin, SpectralFormer, and ConvViT, can achieve sequential representations from the HSI. This produces classification maps of comparable quality to the classical models mentioned earlier. Models based on gating mechanisms (i.e., gMLP) can extract contextual information of HSI data in both the spectral and spatial domains due to their ability to process data recursively, thus enabling relatively high classification results.

In the enlarged part displayed in the red box (see Fig. 5), on the IP dataset, the region shapes of several land covers [i.e., *Alfalfa* (C1), *Corn-Mintill* (C3), *Grass-Pasture* (C5), *Soybean-Notill* (C10), and *Soybean-Clean* (C12)] are narrow and irregular; the sample distributions are also relatively discrete. Owing to the

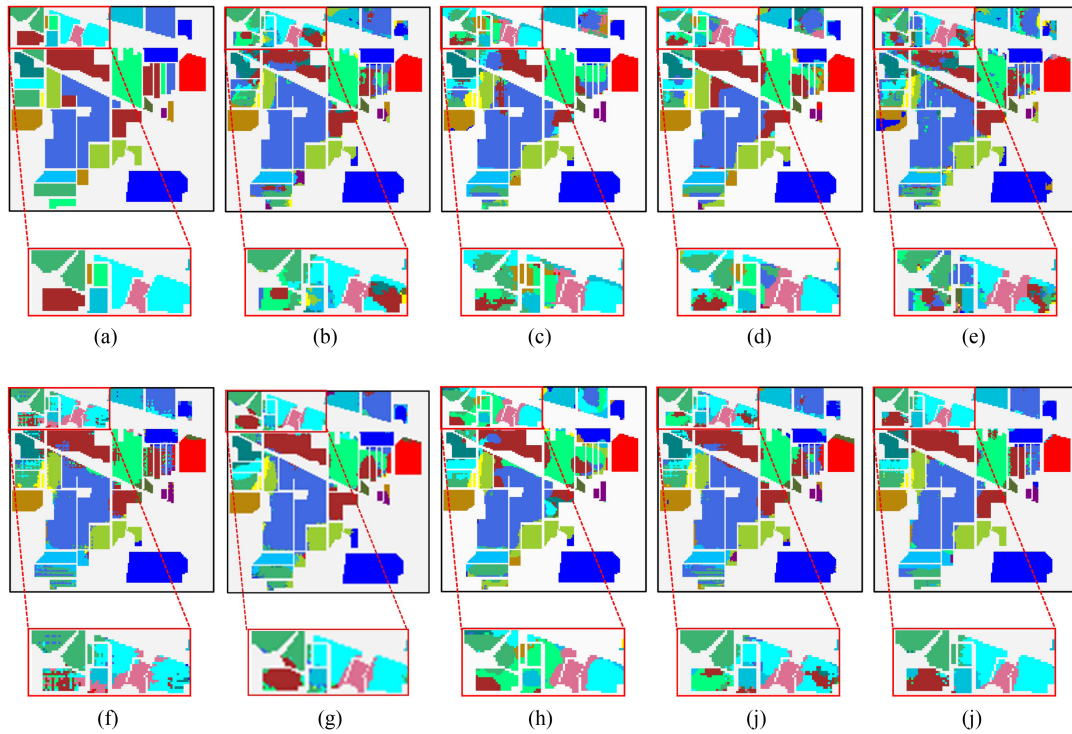


Fig. 5. Classification maps produced by different models for the IP dataset with 1% training samples. (a) Ground truth. (b) CNN (86.06%). (c) GhostNet (77.61%). (d) MAGCaps (82.37%). (e) ViT (82.05%). (f) Swin (79.99%). (g) SpectralFormer (84.77%). (h) ConvViT (78.67%). (i) gMLP (85.27%). (j) **S²MoINet (87.79%)**.

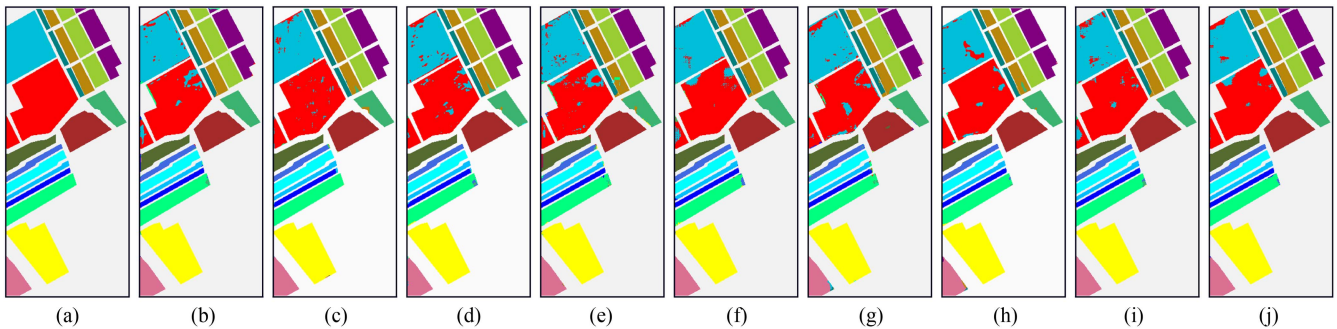


Fig. 6. Classification maps produced by different models for the SV dataset with 1% training samples. (a) Ground truth. (b) CNN (98.83%). (c) GhostNet (98.59%). (d) MAGCaps (97.28%). (e) ViT (98.25%). (f) Swin (98.76%). (g) SpectralFormer (97.33%). (h) ConvViT (97.62%). (i) gMLP (98.66%). (j) **S²MoINet (98.85%)**.

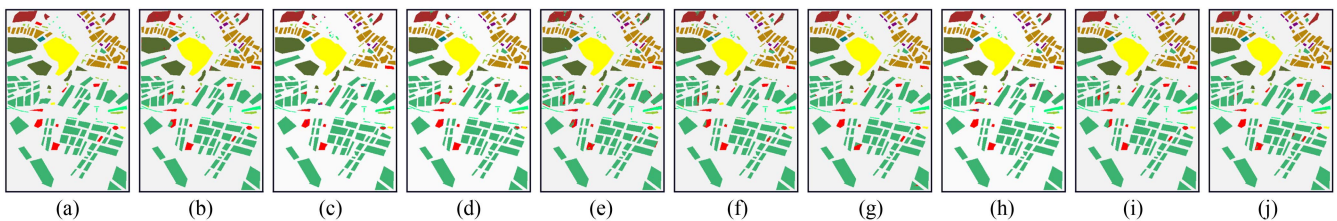


Fig. 7. Classification maps produced by different models for the TF dataset with 1% training samples. (a) Ground truth. (b) CNN (99.22%). (c) GhostNet (98.40%). (d) MAGCaps (98.21%). (e) ViT (98.94%). (f) Swin (98.86%). (g) SpectralFormer (98.85%). (h) ConvViT (97.17%). (i) gMLP (99.04%). (j) **S²MoINet (99.27%)**.

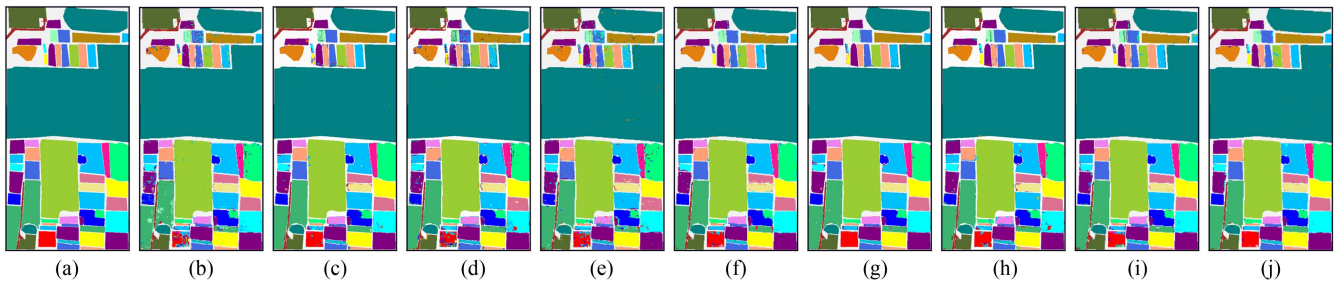


Fig. 8. Classification maps produced by different models for the HH dataset by different classifier. (a) Ground truth. (b) CNN (97.42%). (c) GhostNet (97.92%). (d) MAGCaps (96.01%). (e) ViT (96.56%). (f) Swin (96.86%). (g) SpectralFormer (97.97%). (h) ConvViT (97.56%). (i) gMLP (97.52%). (j) S^2 MoINet (98.04%).

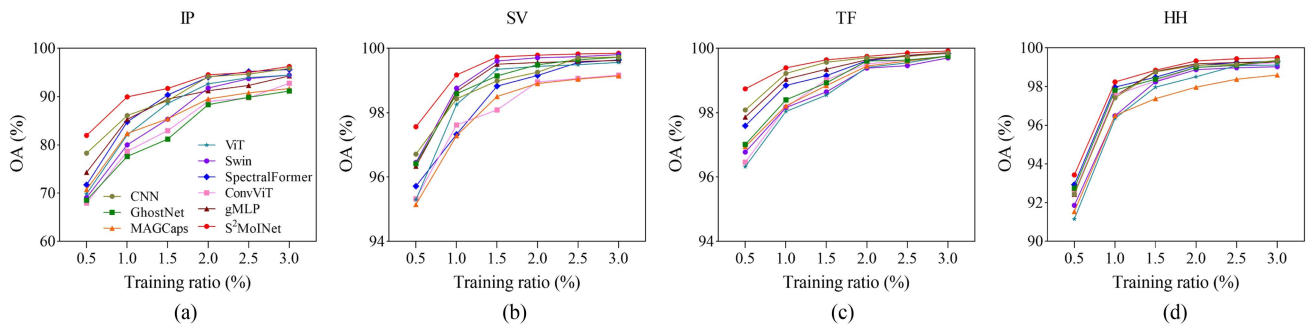


Fig. 9. Classification performance of different models at different training ratios for each HSI dataset with 15×15 input scale. (a) IP. (b) SV. (c) TF. (d) HH.

narrow shape of the *Grass-Pasture* (C5) region, which is in proximity to the *Corn-Mintill* (C3) region, many deep learning models struggle with feature extraction and classification. This brings the misidentification of most ground objects in the *Grass-Pasture* (C5) region as *Corn-Mintill* (C3). In addition, the regions labeled *Soybean-Notill* (C10) and *Soybean-Clean* (C12) in the IP dataset contained samples with spatial proximity. These regions represented the same crop but in different maturity stages, resulting from similarities in their characteristics with few differences. However, previous deep learning models are influenced by each other's structures while learning their unique features. Consequently, the models may exhibit mutual misclassification due to poor feature discriminability.

Similarly, the classification results for the HH dataset obtained by the proposed S^2 MoINet model are also presented in Fig. 8. When comparing the classification performance of different deep learning models, the MAGCaps model has the poorest performance with an accuracy of 96.01%, followed by the ViT and Swin model with 96.56% and 96.86%, while the classification accuracy of other compared models is all above 97%. In contrast, the S^2 MoINet model achieves the best classification accuracy with a rate of 98.04%. This result demonstrates the efficacy of our proposed model for HSI classification task. Besides, it can also be seen that the experimental results on each HSI dataset of S^2 MoINet are all optimal, which also effectively reflects the generalization of our proposed model for HSI classification.

The S^2 MoINet model's classification map in this region displays a relatively smooth outcome thanks to its capability of acquiring multiorder spatial-spectral interaction characteristics of the HSI. However, it is worth noting that the misclassified

pixels by the S^2 MoINet mostly occur at the edges of the region, indicating the difficulty in accurately classifying pixels in these areas due to the smoothing effect of depth-wise convolution. Further research is required to effectively leverage the spatial information inherent in the HSI.

In addition, we also compare the changes in the classification accuracy of different models as the number of training samples increases, as shown in Fig. 9. The results indicate that the classification performance of most models improves as the number of training samples increases. In contrast, the S^2 MoINet consistently maintains high accuracy for all the datasets, and its classification performance steadily improves with an increase in training samples, which also fully illustrates the generalization and stability of our proposed model. Compared with other models, the S^2 MoINet model outperforms them by achieving the best classification results with a training ratio of only 1%. Nonetheless, we have noticed a gradual decline in the rate of model performance optimization as the number of training samples increases, which means that the performance of the model is less sensitive to changes in training samples. This indicates that our proposed model is particularly well suited for classifying HSI data with small sample sizes.

D. Model Analysis

In addition to the learnable parameters within networks and hyperparameters required in the training process, the settings of certain model parameters are also essential for achieving optimal classification performance. Therefore, it is crucial to investigate proper parameter settings. We conducted an analysis of several

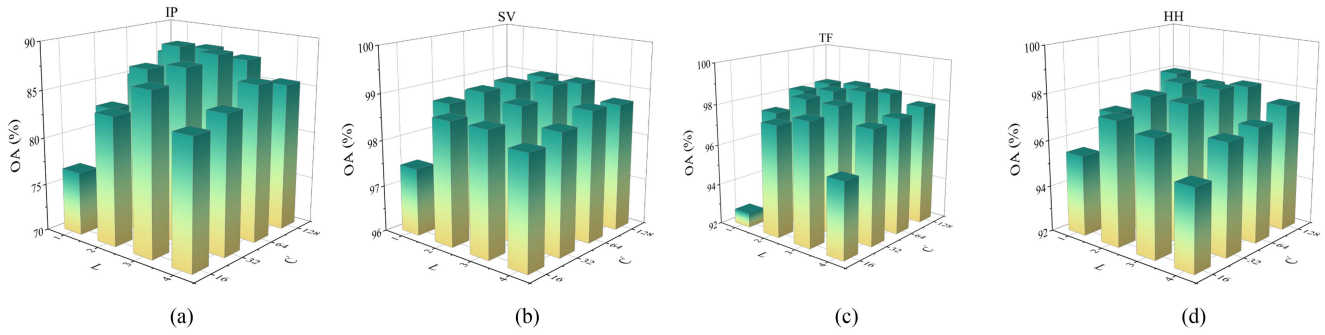


Fig. 10. Classification performance with different L and C' settings of the S²MoINet model on various datasets. (a) IP. (b) SV. (c) TF. (d) HH. L represents the number of MoS²I blocks, and C' represents the number of channels input to the MoS²GM module.

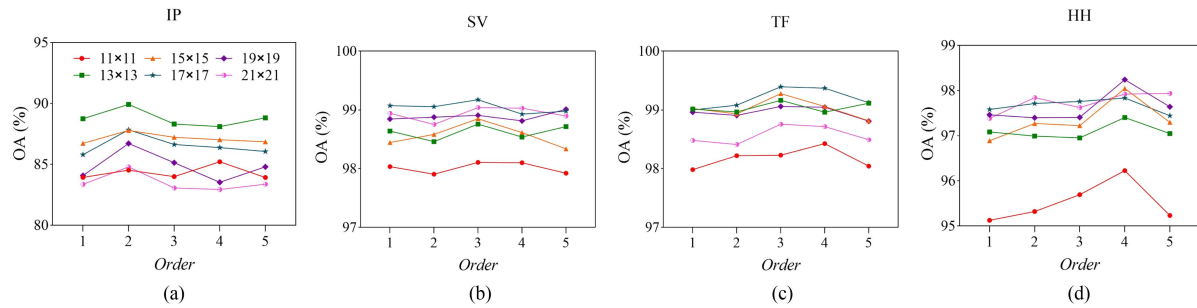


Fig. 11. Classification performance of the S²MoINet model for each HSI dataset at different *Scale* and *Order* with 1% training ratio. (a) IP. (b) SV. (c) TF. (d) HH.

parameters that have an impact on both the classification performance and the training process, which include: 1) the number of MoS²I blocks (L), which can affect the ability of multiorder interactive information extraction; 2) the number of channels input to the MoS²GM module (C'), which can affect the ability of deep spatial–spectral feature extraction; and 3) *order* and *scale*: they represent the order of the S²MoINet model, which can better represent the semantic information of different levels of HSI, and the scale of HSI input data, which can realize the effective representation of the ground object information with different granularity in the HSI, respectively.

1) *Parameter Sensitivity Analysis*: To assess the efficacy of the proposed MoS²I block, we examined the classification performance of each dataset using a 1% training ratio, focusing on the parameters L and C' , as depicted in Fig. 10. We analyzed the influence of various L and C' values on the network, which are selected from 1, 2, 3, 4 and 16, 32, 64, 128, respectively. A comprehensive analysis of the classification results under these parameter settings reveals that increasing C' significantly improves classification accuracy. However, when L exceeds a certain threshold, performance stabilizes and then decreases with further increases in L . This decline is attributed to overfitting caused by excessive L values, which adversely affects classification performance. As the C' parameter increases, the model's classification performance follows a similar trend to that of L —first increasing and then decreasing. Moreover, for all the experimental datasets, the overall classification accuracy generally exhibits a rising and then falling trend as L and C' change. Consequently, to strike a balance between time consumption

TABLE VII
CLASSIFICATION RESULTS OF THE ABLATION STUDY FOR EACH HSI DATASET OF 15×15 INPUT SCALE AT 1% TRAINING RATIO

No.	Variants	S ² FRF	MoS ² I	IP	SV	TF	HH
V1	w/o MoS ² I	✓	✗	85.73	96.78	95.13	90.78
V2	with 1oS ² I	✗	✓	85.87	98.71	98.75	97.57
V3	with 2oS ² I	✗	✓	<u>87.41</u>	98.51	98.59	97.61
V4	with 3oS ² I	✗	✓	87.09	<u>98.73</u>	<u>98.92</u>	97.70
V5	with 4oS ² I	✗	✓	87.13	98.43	98.69	<u>97.78</u>
V6	with 5oS ² I	✗	✓	86.74	98.68	98.62	97.58
V7	S ² MoINet	✓	✓	87.79	98.85	99.27	98.04

* The top two values in each column are emphasized by bold and underline formatting, denoting the highest and second-highest results.

and classification performance, we set the parameter L to 3 and C' to 64 for subsequent specific research and analysis of the S²MoINet model on each dataset.

2) *Performance Analysis*: To specifically investigate the S²MoINet model at different orders (i.e., *Order* = 1, 2, 3, 4, 5) and input scales (i.e., *Scale* = 11×11, 13×13, 15×15, 17×17, 19×19, 21×21), we further compare and analyze the classification performance on each HSI dataset with its OA at a fixed training ratio of 1%. The specific results are shown in Fig. 11.

Generally, the multiorder spectral–spatial information (i.e., abstract features) of the HSI ground objects considered in the model gradually aggrandizes with the increase of the feature order extracted by the model, thus resulting in a certain degree of improvement in the classification performance of the multiorder

TABLE VIII
PARAMETERS OF NINE METHODS ON FOUR DATASETS

Dataset	CNN	GhostNet	MAGCaps	ViT	Swin	SpectralFormer	ConvViT	gMLP	S ² MoINet
IP	3.28	0.06	0.07	4.21	4.22	4.15	2.89	4.37	4.03
SV	3.32	0.06	0.07	4.23	4.24	4.17	2.91	4.38	4.06
TF	2.92	0.05	0.06	3.98	4.01	3.85	2.73	4.01	3.81
HH	3.45	0.07	0.09	4.35	4.39	4.21	2.98	4.52	4.13

model. According to the results in Fig. 11, it can be found that owing to the different resolutions of the HSI datasets, the appropriate input scale and the order of feature extraction also vary. The IP dataset has a relatively low resolution of 20 mpp, which means that it contains less information in each pixel, and most of the detailed information will be lost when utilizing an excessively large input scale for feature extraction, thus reducing the classification performance of the model. As a result, we achieved the highest classification accuracy of 89.93% for the IP dataset when utilizing a smaller input scale of 13×13 and second-order S²MoINet model. For the SV and TF datasets, their resolutions are relatively similar to each other (i.e., 3.7 and 2.25 mpp), and the resolution is somewhat higher than that of the IP dataset, and the amount and complexity of the information contained in the pixels have also increased. Therefore, when processing the HSI ground objects, a larger input scale and feature extraction order are required. The optimal classification accuracy is obtained when the input scale is 17×17 and the order of S²MoINet model is 3, which are 99.39% (SV) and 98.24% (TF), respectively. The HH dataset has the highest resolution of 0.043 mpp, and each pixel in the data has more abundant and complex information, so it can obtain sufficient feature information when the input scale is 19×19 and the order of the S² MoINet model is 4 and finally obtain the highest classification accuracy of 98.24%.

These results demonstrate that for high-resolution datasets with complex structural features, the proposed S²MoINet model can better extract abstract multiorder features, enabling a more comprehensive description of HSI data features. In low-resolution HSI datasets with simple feature structures, the model can also extract corresponding low level yet effective feature representations. Consequently, the proposed model can better describe the heterogeneous structural information of HSI objects at different levels.

3) *Ablation Experiment*: To confirm the efficacy of the S²FRF and MoS²I blocks in acquiring and classifying HSI features effectively, we conducted a series of ablation experiments for each dataset using a 15×15 input scale at a 1% training ratio (see Table VII). Specifically, our experiments have seven variants (from V1 to V7); the specific settings are as follows.

- 1) V1: the model only has the S²FRF block, and without the MoS²I block, i.e., no multiorder spectral–spatial interaction, we named it as **w/o MoS²I**.
- 2) V2–V6: there is no S²FRF block, and only the MoS²I block is used to consider the k th-order ($k=1,2,3,4,5$) information of HSI data, i.e., only k th-order spectral–spatial interaction is performed, we named it as **with koS^2I** .
- 3) V7: the proposed **S²MoINet** model.

It is evident that the classification accuracy of V1 is relatively the lowest when only simple spatial–spectral feature extraction is performed by the S²FRF block. In contrast, the classification accuracy of V2–V6 improves when considering only the multiorder feature interaction of the data due to the increased features. These variants achieve the second-best classification effect at different orders, depending on the dataset.

Given the varying resolutions of each dataset, the complexity of information contained in each pixel also differs. Consequently, the IP dataset achieves better results using the low-order MoS²I block (i.e., second-order), whereas the SV and TF datasets exhibit the relatively superior classification performance when the MoS²I block order is set to 3. The HH dataset, with more complex and detailed information and a stronger nonlinear structure of its ground object, attains better classification results when the MoS²I block order is set to 4.

Compared to other variants, V7 (S²MoINet) achieves the optimal classification performance, providing clear the evidence of the effectiveness of the two proposed blocks: S²FRF and MoS²I, for HSI feature extraction and classification. Both the blocks are indispensable; removing the S²FRF block or disabling the MoS²I block will degrade the model’s classification accuracy.

V. DISCUSSION

In this section, another metric is utilized to evaluate the performances of the proposed S²MoINet and several comparative models. The model’s computational complexity can be assessed by comparing the number of parameters for different models; the specific parameters of the S²MoINet and other comparative models on four HSI datasets are shown in Table VIII. From the table, we can observe that the convolution variants models (such as GhostNet and MAGCaps) show the best and second-best results in terms of parameter quantity due to the characteristics of their lightweight design. The traditional CNN model, although its classification performance is mostly suboptimal in relative terms, has the largest number of model parameters. There are also a lot of parameters in the gMLP model and the transformer variation models (such as, ViT, Swin, SpectralFormer, and ConvViT).

In contrast, our proposed S²MoINet model incurs an increase in the number of parameters compared with the other models (ranking fifth highest) due to its ability to consider multiorder information of HSI data. Nevertheless, our model exhibits the best classification performance among all the models evaluated, indicating that it achieves a better balance between classification accuracy and model complexity.

VI. CONCLUSION

In this article, we designed a novel S²MoINet to settle the problem that previous models do not consider the MoS²I features of the HSI. Our approach placed greater emphasis on the multiorder interactions of HSI, which enabled it to effectively depict local detailed spectral differences and convey multiorder spatial interaction information during feature extraction. The proposed model demonstrated success in extracting discriminative features for HSI classification tasks, especially in processing the edges of noise class regions and pixels in uniform regions of mixed ground objects in HSI datasets containing high-resolution complex information. The proposed S²MoINet outperformed both the traditional and cutting-edge deep learning techniques, according to extensive experiments on four popular HSI datasets.

In future work, we will enhance the model by incorporating advanced techniques to increase its applicability for HSI classification tasks and developing a lightweight network to reduce computational complexity without compromising classification performance. Furthermore, we aim to investigate the impact of the proposed model under varying conditions to obtain MoS²I features of HSI data, which would make the model more interpretable.

REFERENCES

- [1] 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.
- [2] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 482.
- [3] J. Feng, G. Bai, D. Li, X. Zhang, R. Shang, and L. Jiao, "MR-selection: A meta-reinforcement learning approach for zero-shot hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500320.
- [4] J. Feng, N. Zhao, R. Shang, X. Zhang, and L. Jiao, "Self-supervised divide-and-conquer generative adversarial network for classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536517.
- [5] J. Yuan, S. Wang, C. Wu, and Y. Xu, "Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: A case study of Wuhan," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3972–3991, 2022.
- [6] M. Shimoni, R. Haelterman, and C. Perneel, "Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.
- [7] L. Ravikanth, D. S. Jayas, N. D. White, P. G. Fields, and D.-W. Sun, "Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products," *Food Bioprocess Technol.*, vol. 10, pp. 1–33, 2017.
- [8] S. Wang et al., "Cross-scale sensing of field-level crop residue cover: Integrating field photos, airborne hyperspectral imaging, and satellite data," *Remote Sens. Environ.*, vol. 285, 2023, Art. no. 113366.
- [9] K. Jacq et al., "Sedimentary structure discrimination with hyperspectral imaging in sediment cores," *Sci. Total Environ.*, vol. 817, 2022, Art. no. 152018.
- [10] J. Zhao, H. Yan, and L. Huang, "A joint method of spatial–spectral features and BP neural network for hyperspectral image classification," *Egypt. J. Remote Sens. Space Sci.*, vol. 26, no. 1, pp. 107–115, 2023.
- [11] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognit.*, vol. 103, 2020, Art. no. 107298.
- [12] 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.
- [13] 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.
- [14] H. Zhou, C. Zhang, X. Zhang, and Q. Ma, "Image classification based on quaternion-valued capsule network," *Appl. Intell.*, vol. 53, no. 5, pp. 5587–5606, 2023.
- [15] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.
- [16] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8246–8257, Dec. 2020.
- [17] S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza, "Morphological convolutional neural networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8689–8702, 2021.
- [18] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
- [19] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 884.
- [20] H. Huang, C. Pu, Y. Li, and Y. Duan, "Adaptive residual convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2520–2531, 2020.
- [21] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.
- [22] H. Gao, Y. Miao, X. Cao, and C. Li, "Densely connected multiscale attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2563–2576, 2021.
- [23] Y. Fang, Q. Ye, L. Sun, Y. Zheng, and Z. Wu, "Multi-attention joint convolution feature representation with lightweight transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513814.
- [24] 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.
- [25] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [27] Y. Qing, W. Liu, L. Feng, and W. Gao, "Improved transformer net for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2216.
- [28] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 8671–8685, 2021.
- [29] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5518615.
- [30] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5514715.
- [31] F. Zhou, R. Hang, J. Li, X. Zhang, and C. Xu, "Spectral-spatial correlation exploration for hyperspectral image classification via self-mutual attention network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6010205.
- [32] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
- [33] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
- [34] S. Guo, Q. Jin, H. Wang, X. Wang, Y. Wang, and S. Xiang, "Learnable gated convolutional neural network for semantic segmentation in remote-sensing images," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1922.
- [35] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S. N. Lim, and J. Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 10353–10366.

- [36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [37] C. Li et al., "Effective multiscale residual network with high-order feature representation for optical remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 6003105.
- [38] W. Guo, G. Xu, B. Liu, and Y. Wang, "Hyperspectral image classification using CNN-enhanced multi-level Haar wavelet features fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6008805.
- [39] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2016.
- [40] L. Chen, Z. Wei, and Y. Xu, "A lightweight spectral-spatial feature extraction and fusion network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1395.
- [41] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2021.
- [42] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
- [43] X. Li, M. Ding, and A. Pižurica, "Fully group convolutional neural networks for robust spectral-spatial feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5509314.
- [44] Z. Meng, J. Zhang, F. Zhao, H. Liu, and Z. Chang, "Residual dense asymmetric convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 3159–3162.
- [45] F. F. Shahraki and S. Prasad, "Graph convolutional neural networks for hyperspectral data classification," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2018, pp. 968–972.
- [46] J. Chen, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "Automatic graph learning convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520716.
- [47] M. E. Paoletti et al., "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019.
- [48] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 516.
- [49] L. Zhang, Y. Zeng, J. Zhao, and J. Lan, "A novel global-local block spatial-spectral fusion attention model for hyperspectral image classification," *Remote Sens. Lett.*, vol. 13, no. 4, pp. 343–351, 2022.
- [50] X. Xue, H. Zhang, B. Fang, Z. Bai, and Y. Li, "Grafting transformer on automatically designed convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531116.
- [51] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.
- [52] W. Zhou, S. -i. Kamata, Z. Luo, and H. Wang, "Multiscanning strategy-based recurrent neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5521018.
- [53] Z. Hua, X. Li, J. Jiang, and L. Zhao, "Gated autoencoder network for spectral-spatial hyperspectral unmixing," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3147.
- [54] Q. Meng, M. Zhao, L. Zhang, W. Shi, C. Su, and L. Bruzzone, "Multilayer feature fusion network with spatial attention and gated mechanism for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6510105.
- [55] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 127–143, 1993.
- [56] R. O. Green et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, 1998.
- [57] X. Zhang, B. Zhang, L. Zhang, and Y. Sun, "Hyperspectral remote sensing dataset for tea farm [J/DB/OL]," *Digit. J. Glob. Change Data Repository*, 2017, doi: [10.3974/geodb.2017.03.04.V1](https://doi.org/10.3974/geodb.2017.03.04.V1).
- [58] 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, 2020, Art. no. 112012.
- [59] M. E. Paoletti, J. M. Haut, N. S. Pereira, J. Plaza, and A. Plaza, "GhostNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10378–10393, Dec. 2021.
- [60] M. E. Paoletti, S. Moreno-Álvarez, and J. M. Haut, "Multiple attention-guided capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5520420.
- [61] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [62] Z. Zhao, D. Hu, H. Wang, and X. Yu, "Convolutional transformer network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6009005.
- [63] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9204–9215.



Yanan Jiang (Student Member, IEEE) received the B.S. degree in mathematics and applied mathematics from the Hebei University of Science and Technology, Shijiazhuang, China, in 2017, and the M.S. degree in mathematics from the China University of Geosciences, Beijing, China, in 2020. She is currently working toward the Ph.D. degree in applied mathematics with Beijing Normal University, Beijing.

Her research interests include applied mathematics, machine learning, statistical learning, pattern recognition, and computer vision.



Heng Zhou (Student Member, IEEE) received the B.Eng. degree in electronic and information science from the Beijing University of Posts and Telecommunications, Beijing, China, in 2017, and the M.Eng. degree in computer technology from the China University of Geosciences, Beijing, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with China Agricultural University, Beijing.

His research interests include deep learning, computer vision, and pattern recognition.



Zitong Zhang (Student Member, IEEE) received the B.S. and M.S. degrees in mathematics in 2017 and 2020, respectively, from the China University of Geosciences, Beijing, China, where she is currently working toward the Ph.D. degree in earth exploration and information technology with the School of Earth Sciences and Resources.

Her research interests include deep learning and hyperspectral image classification.



Chunlei Zhang received the M.Eng. degree in coal petrology and coalfield geology from the Taiyuan University of Technology, Taiyuan, China, in 1997, and the Ph.D. degree in mineral resource prospecting and exploration from the China University of Petroleum, Beijing, China, in 2000.

He majored in geostatistics, reservoir characterization and reservoir engineering. From 2002 to 2004, he was a Postdoctoral Researcher with the China University of Petroleum. Since 2004, he has been with Beijing Zhongdi Runde Petroleum Technology Co., Ltd., Beijing, China. His research interests include geostatistics, pattern recognition, machine learning, and computer vision.



Kai Zhang received the B.S. degree in information science and technology from Zhengzhou University, Zhengzhou, China, in 2018, and the M.Eng. degree in computer technology from the China University of Geosciences, Beijing, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with China Agricultural University, Beijing.

His research interests include deep learning and computer vision.