scientific reports

OPEN

Spectral-spatial wave and frequency interactive transformer for hyperspectral image classification

Tahir Arshad[1], Bo peng[1], Ali Rahman[2], Rahim khan[3], Sajid Ullah khan[4]✉, Sultan Alnazi[4] & Nazik Alturki[5]

Efficient extraction of spectral-spatial features is essential for accurate hyperspectral image (HSI) classification, where capturing both local texture and global semantic relationships is critical. While Convolutional Neural Networks (CNNs) and Transformers have shown strong capabilities in modeling local and global dependencies, most existing architectures operate directly on raw spectral-spatial inputs and lack explicit mechanisms for frequency-domain decomposition thereby overlooking potentially discriminative phase and frequency components. To address this limitation, we propose a Spectral-Spatial Wave and Frequency Interactive Transformer for HSI classification, which integrates frequency-aware and phase-aware token representations into a unified Transformer framework. Specifically, our model first employs a CNN backbone to extract shallow spectral-spatial features. These are then processed by a novel Frequency Domain Transformer Encoder, composed of two complementary branches: (i) a Spectral-Spatial Frequency Generator that extracts multiscale frequency features, and (ii) a Spectral-Spatial Wave Generator that encodes phase and amplitude characteristics as complex-valued wave tokens. A Spectral-Spatial Interaction Module fuses these components, followed by a Local-Global Modulator that refines semantic representations from multiple perspectives. Extensive experiments on five benchmark HSI datasets, demonstrate the effectiveness of our approach. The proposed model achieves state-of-the-art classification performance, with Overall Accuracies of 98.49%, 98.60%, 99.07%, 98.29%, and 97.97%, consistently outperforming existing methods.

Keywords Attention module, Convolutional neural network, Hyperspectral image classification, Frequency domain, Vision transformer

The hyperspectral image data has numerous narrow bands containing a substantial amount of information. HSI is a part of earth observation[1] extensively used in agriculture[2], mineralogy[3], and environmental sciences. HSI data captures spectral information from numerous contiguous spectral bands of surface objects. Numerous approaches have been proposed in the last decade to address the challenges in HSI. This work proposes to solve the problem of feature extracting in the frequency domain. Initially, researchers devised traditional methods of machine learning for the categorization of HSI, such as k-nearest neighbor[4], logistic regression[5], Bayesian estimation[6], and support vector machines[7]. Still, it has been noted that these traditional categorization models frequently lead to misclassification. Furthermore, other techniques have been devised to reduce dimensionality and retrieve spectrum information, including principal component analysis (PCA)[8] and linear discriminative analysis (LDA)[9].

However, these techniques often fail to consider the spatial correlation between pixels in a spatial dimension, which is essential for achieving effective spatial feature extraction. In order to tackle this problem, several mathematical operators have been devised, including the morphological profile[10,11]. Deep learning is well

[1]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China. [2]School of Civil Engineering, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, UK. [3]College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China. [4]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Alkharj, Kingdom of Saudi Arabia. [5]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia. ✉email: sk.khan@psau.edu.sa

| https://doi.org/10.1038/s41598-025-12489-3

recognized and highly successful technique for extracting features in the context of hyperspectral classification[12]. For instance, a stacked autoencoder (SAE)[13], a deep belief network (DBN)[14] involves the extraction of depth and invariant characteristics from hyperspectral data. The obtained characteristics are then utilized in logistic regression to tackle the challenge of hyperspectral image categorization. However, it is crucial to acknowledge that both SAE and DBN encode spatial information as vectors during the pretraining phase, leading to the unavoidable loss of spatial information.

It is important to mention that a CNN[15–17]. In[18,19] introduced a HybridSN network to simplify the net framework by integrating spectral-spatial information using 3D and 2D convolutions. The basic hybrid model outperforms both the 3D-CNN model and the 2D-CNN model in terms of computing efficiency, and it demonstrates superior performance in handling the issue of small sample size. A spectral–spatial capsule network proposed by Paoletti et al.[20] to reduce the complexity of convolutional networks. some recent works[21,22] investigate distinctive methods to enhance the discrimination of spectral–spatial.

features. Utilizing the attribute profiles Aptoula et al.[23] to provide the spectral and geometric features of HSI for the CNN. To address the issue of inadequately labelled samples, Li et al.[21] proposed a data augmentation technique. Chen et al.[22] proposed to combine a CNN with Gabor filters to enhance texture and edge information. Mei et al.[24] utilized a CNN to identify sensor-specific features for HSI. Li et al.[25] improved spectral spatial features by utilizing a PCA and a fully convolutional network. The attention mechanism is inspired by the human visual system to identify important regions from images for classification[26]. The attention recurrent convolutional network becomes excessively complex when combining long short term memory (LSTM) and CNN to extract attention features. a two branch attention module proposed by Haut et al.[26] where one branch is utilized to generate an attention mask and the other branch is utilized to extract convolutional features. Then, by multiplying the attention mask with the convolutional attention features are obtained. Different from earlier methods an attention modules and interactive feature enhancement module that integrate into a simple component. In order to extract attention features, the suggested feature enhancement module takes use of the correlation between the hyperspectral pixels inside an HSI cube.

Recently vision transformers (ViT)[27] architecture is based on self-attention mechanisms combined with multilayer perceptron's (MLPs), which excel at capturing long-range dependencies between tokens in a sequence. Originally designed for machine translation tasks in natural language processing[28], vision transformers have gained significant attention due to their remarkable success, prompting researchers to explore their application in image processing and computer vision. In contrast to convolutional networks, transformers offer a fresh perspective and new possibilities for improving performance at the attention level. The key factor behind the transformer's success lies in its ability to effectively model long-range relationships, overcoming limitations faced by traditional architectures. The transformer has been recently introduced in HSI. Hong et al.[29] and He et al.[30] treated HSI patches as sequential data from the spectral dimension, utilizing group-spectral embedding and a transformer encoder module to capture locally detailed spectral representations. Roy et al.[31]suggested An attention-based adaptive spectral–spatial kernel-improved residual network (A2S2K-ResNet) was proposed to adaptively select 3D convolutional kernels for effective extraction of salient spectral–spatial information, achieving superior performance compared to other methods on several HSI datasets. Roy et al.[32] The attention mechanism was integrated with morphological operations to design the input patch for the Transformer, enriching the spectral–spatial information by incorporating morphological features and thereby enhancing classification performance. Zhao et al.[33] designed lightweight network model called GSCViT was designed, which integrates groupwise separable convolution into the Transformer framework to effectively extract both local and global spatial features. Z. Meng et al.[34]introduced A Global–Local Multi-Granularity Transformer was proposed to capture both local and global features by employing a Multi-Granularity Spatial Feature Extraction (MGAFE) block for comprehensive spatial information extraction at various granularities, and a Multi-Granularity Spectral Feature Extraction (MGEFE) block to effectively leverage spectral information across multiple scales. Meng et al.[35] introduced a spectral spatial MLP architecture to capture long range dependencies and enhanced the classification results. Fan et al.[36] proposed a novel frequency topology interaction network to capture both spatial and frequency domain features This approach combines CNNs and self-attention mechanisms to capture local and global contextual information, enhancing feature representation and addressing challenges like spectral ambiguity and material heterogeneity. By integrating both global and local information, this approach significantly enhances the model's perceptual capabilities. Sun et al. introduced the Spectral-Spatial Feature Tokenization Transformer (SSFTT)[37], designed to capture both spectral-spatial and high-level semantic features. The model initially extracts shallow spectral and spatial features using a combination of 3D and 2D convolution layers. It then applies a Gaussian-weighted feature tokenizer for feature transformation, with the transformed features being fed into a transformer encoder for representation and learning. Ma et al.[38] proposed a Vision Transformer incorporating a lightweight self-Gaussian attention (LSGA) mechanism to extract global deep semantic features. Yang et al.[39] introduced a hyperspectral image transformer that incorporates two key modules: spectral adaptive 3D convolution projection and convolution permutator, designed to capture subtle spectral and spatial information. Bin Li et al.[40] proposed a multi-granularity vision transformer using a semantic tokens transformer to learn multi-granularity features and enhance accuracy. They employed the LFE module to extract local features. E. Ouyang et al.[41] developed the HybridFormer network, which combines CNNs for extracting shallow features with a spectral-spatial attention (SSA) based transformer encoder for learning semantic features. Transformers designed for HSI classification represent spatial patches or spectral bands as tokens. While these tokens enable the modelling of intricate relationships across spatial and spectral dimensions, the high token count results in substantial computational overhead. However, the aforementioned methods primarily focus on feature extraction from the spectral-spatial domain and do not fully leverage the information available in the frequency domain.

While analysing images in the frequency domain is a well-established and efficient technique for natural images, most existing transformer-based classifiers focus on feature extraction from the spectral-spatial domain and overlook the potential of the frequency domain. This gap motivates the exploration of hyperspectral images from the frequency domain. The model proposed in[42] is designed to extract both high- and low-frequency information. However, their approach, based on the CNN architecture, has limitations in capturing global contextual information. Qiao et al.[43] proposed a novel hierarchical dual-frequency transformer network that captures high-frequency and low-frequency features through a dual-branch structure. Shi et al.[44] proposed a multiscale conv aided Fourier transformer model to extract spectral spatial features in frequency domain. However, the aforementioned methods does not fully exploit the multiscale frequency features. Despite the effectiveness of deep learning methods in HSI classification, most existing approaches either focus on spatial textures or spectral signatures independently, while overlooking the crucial role of frequency- and phase-based representations. However, both spectral frequency variations and spatial phase patterns are essential for distinguishing subtle material differences. To address this gap, we propose a unified framework consisting of four key modules, each designed to extract and integrate distinct aspects of frequency and phase information. To address this issue we proposed a spectral spatial phase frequency domain transformer encoder to fully exploit the multiscale spectral spatial features from HSI data.

The main contribution of this paper is the following:

- The proposed framework is designed for HSI. It utilizes a combination of spectral spatial feature extractor with frequency domain transformer encoder block to extract phase and frequency features.
- We used 3DConv and 2DConv layer to effectively capture the spectral spatial shallow frequency domain features from HSI cube. After that frequency domain transformer encoder block are proposed which consist of spectral spatial wave token generator and spectral spatial multiscale frequency token generator to learn frequency domain features. The Spectral-Spatial Wave Generator models spatial frequency components through phase and amplitude modulation across spatial locations. While, the Multiscale Frequency Token Generator captures spectral frequency variations by decomposing features at multiple spectral resolutions.
- An efficient Spectral-Spatial Interaction Module is introduced to enable the effective fusion of phase, amplitude, and frequency features, enhancing the model's ability to capture complex spectral-spatial relationships. In addition, a Local-Global Modulator Module is proposed, consisting of two parallel branches designed to extract semantic features from multiple perspectives the local branch focuses on fine-grained spatial details, while the global branch captures broader contextual information. Together, these modules contribute to a more comprehensive and robust representation of hyperspectral data.

The rest of the paper is organized as follows: section II explains recent works; section III Proposed Methodology; section IV Dataset and Experiment Evaluation; section V Results and Discussion; and section VII Conclusion.

Related work

CNN based methods with attention-aided CNNs

Researchers have utilized CNNs for hyperspectral image classification due to their ability to automatically extract features, eliminating the need for complex image preprocessing. Compared to traditional methods, CNN-based approaches deliver superior classification performance. Based on the feature dimensions they process, CNN-based methods can be classified into three categories: 1D-CNN, 2D-CNN, and 3D-CNN. In the early stages of HSI classification, 1D-CNNs were primarily utilized to extract feature vectors along the spectral dimension. For instance, Hu et al.[15]proposed an HSI classification model employing a 1D-CNN with five convolutional layers. Compared to 1D-CNNs, 2D-CNNs offer the advantage of capturing spatial context, effectively modeling spatial dependencies within HSI data. In[11,45], 2D-CNN models were developed to extract spatial features, following a dimensionality reduction step using Principal Component Analysis. Additionally, in[46], a dual-branch network was introduced, integrating both 2D-CNN and 1D-CNN architectures, where the 2D-CNN focused on spatial feature extraction while the 1D-CNN specialized in spectral feature extraction for enhanced HSI classification performance. To better utilize the intrinsic characteristics of HSI data, joint extraction of spatial and spectral information, rather than separate processing, is considered a more effective approach. Consequently, 3D-CNNs have been widely adopted for HSI classification. Chen et al.[47] introduced a 3D-CNN framework incorporating virtual sample enhancement and regularization techniques, achieving promising classification results. Furthermore, inspired by the pyramidal hierarchical structures commonly used in CNNs, Rao et al.[48] proposed the 3D Adaptive Spatial-Spectral Pyramidal Layer CNN model (ASSP-SCNN). This model leveraged multiscale samples during training to fully exploit spatial-spectral features while mitigating overfitting, thereby enhancing classification performance. Xu et al.[46] were the first to propose feeding the entire image directly into the model, rather than inputting pixel patches one at a time, to learn global representations. Building on this idea, Wang et al.[49] introduced FullyContNets, which also utilizes the full image and adaptively aggregates multiple features through a pyramid multi-scale structure. Experimental results demonstrated that leveraging the rich spatial information from the complete image can lead to excellent HSI classification performance. To further assist the CNN in learning better features, the attention mechanism is applied as an enhancement unit to the CNN-based HSI classification. Attention mechanism used in different fields like remote sensing[50], medical image analysis[51], image enhancement[52] and natural language processing[53].

Transformer based and related token generation

Recent breakthroughs in computer vision have been predominantly driven by deep neural networks, though recent work demonstrates that hybrid approaches incorporating classical methods like fast Fourier transform (FFT) can achieve superior spectral sensitivity while maintaining spatial awareness. Recent work[54] introduced

Fast Fourier Convolution (FFC), a novel neural operator that synergistically integrates FFT with standard convolutions. This architecture enables simultaneous non-local feature propagation and multi-scale information fusion through a dual-branch design. In[55] proposed in frequency-domain and spatial-domain feature fusion network (FSFF-Net) for HSI classification, which reduces computational complexity while capturing global features. He et al.[56], proposed frequency domain network with multiscale learnable convolution attention to capture global spatial and frequency domain features. For instance[57], applies FFT repeatedly within non-hierarchical frameworks, resulting in redundant computational complexity without efficiently transitioning from shallow to deep feature representations. Hao et al.[44], proposed MHCFormer for HSI classification which is CNN based transformer module uses fouriormixer module to capture long range dependencies.

While attention mechanisms serve as the predominant token mixer in standard transformer architectures, recent research has explored alternative mixing operations that offer complementary advantages. Tolstikhin et al.[58] demonstrated that spatial MLPs can effectively serve as standalone token mixers, with their MLP-Mixer architecture achieving competitive performance compared to attention-based transformers. Subsequent research has advanced MLP-based architectures through both efficiency improvements and novel token-mixing strategies:[59,60] enhanced MLP models via data-efficient training protocols and redesigned MLP modules, while alternative approaches replaced attention with spectral operations FNet[61] employed Fourier transforms as token mixers for NLP, Global Filter Networks[62] implemented depthwise global convolutions through learnable Fourier filters, and AFNO[63] introduced adaptive Fourier neural operators for dynamic frequency-domain mixing. These developments collectively demonstrate the viability of non-attention token mixing across domains. While these architectures demonstrate strong performance in conventional computer vision tasks, their direct applicability to hyperspectral image classification remains suboptimal due to three domain-specific constraints: spectral spatial tradeoff, high dimensionality and limited training data. However the proposed model is relatively fixed in frequency domain to capture spectral spatial features from frequency domain.

Proposed methodology

Figure 1 depicts the proposed model, this section will elucidate the model's structure and it's functioning. The feature pre-processing stage begins by feeding HSI patches into 3D convolutional layer and 2D convolutional layer for low-level feature extraction after applying Feature Augmentation. Shallow feature maps generated by these layers are then processed using pixel operations to tokenize the features before passing them into the

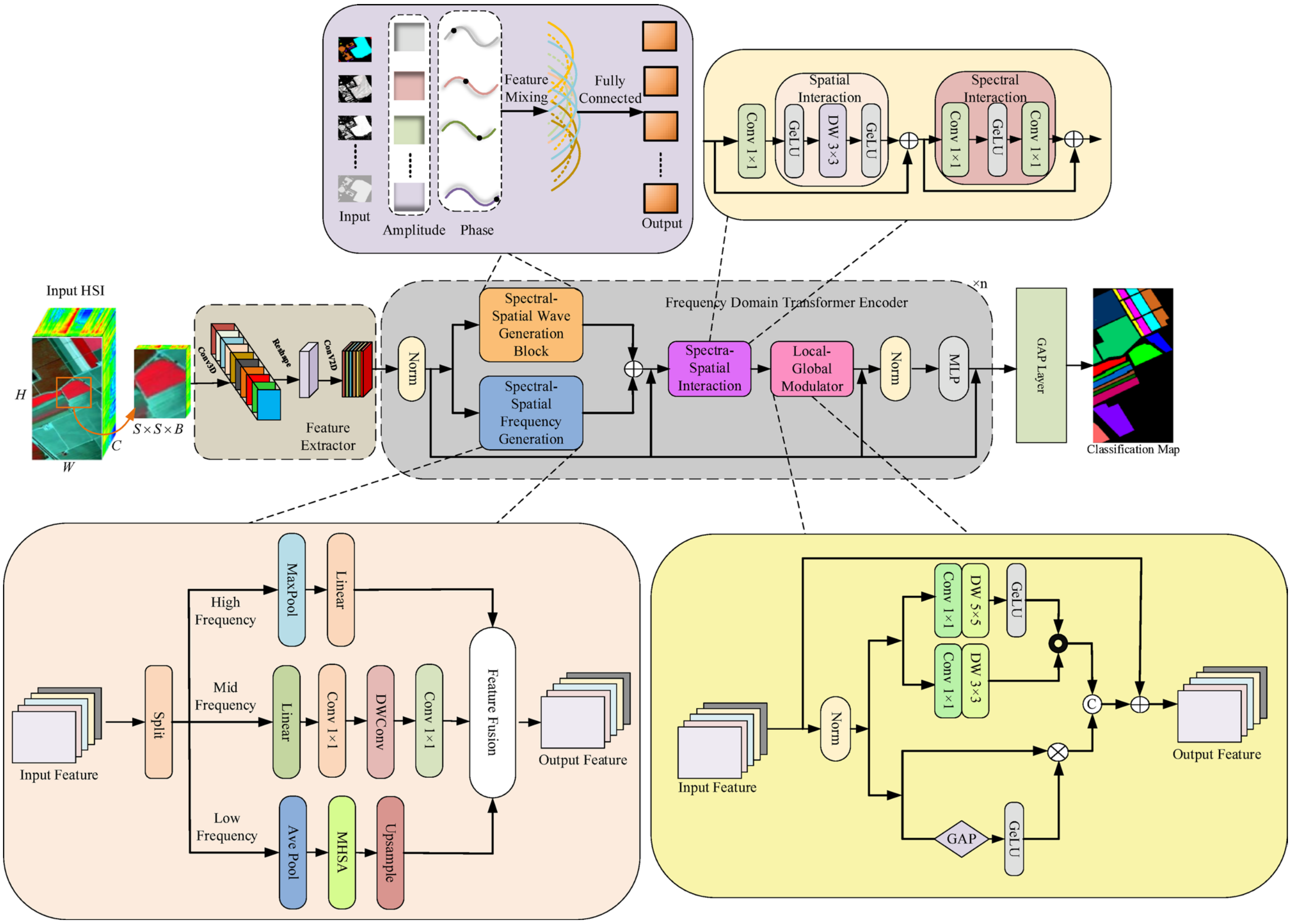


Fig. 1. Illustration of the Proposed Model for HSI classification.

transformer encoder Blocks. The core of the model lies in the transformer Blocks, which are repeated N times to enable interactions between tokens representing different spectral spatial locations in the frequency domain. Finally, a classification head, consisting of a Global Average Pooling (GAP) layer is employed to assign a label to each pixel.

Frequency domain token generation

HSI contain a large number of spectral bands, providing rich spectral information. However, this also results in redundant features and significant computational overhead. As a result, Principal Component Analysis is commonly employed as a dimensionality reduction technique to compress HSI data. To mitigate the risk of overlooking subtle spectral or spatial variations during early representation learning, we propose a dual-branch token generation strategy. The Spectral-Spatial Multiscale Frequency Token Generator extracts broad frequency features at multiple spectral resolutions, capturing both coarse and fine-grained spectral structures across bands. In parallel, the Spectral-Spatial Wave Token Generator models tokens as learnable complex-valued waves, where the amplitude encodes spectral energy distribution and the phase captures localized spatial transitions and material-specific structural patterns. By combining these two complementary mechanisms, our approach enables comprehensive modeling of both global frequency characteristics and local spatial dynamics, enhancing the expressiveness and robustness of hyperspectral feature representations. Let the HSI data be represented by a matrix $X_{hsi} \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ indicate the height and width of the data, and *C* represents the spectral dimension. After applying PCA on the spectral dimension, the HSI data can be represented as$X_{hsi} \in \mathbb{R}^{H \times W \times B}$ suppose thatcontains *N* labelled pixel is a vector$X = \{x_1, x_2 \ldots, x_N\} \in \mathbb{R}^{1 \times 1 \times B}$and their corresponding one hot label vector $Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^{1 \times 1 \times K}$where *K* represents the number of classes. The spatial size$S \times S$around center pixel can be defined as a spectral spatial vector. We begin by utilizing a 3D convolutional layer (3D Conv) and a 2D convolutional layer (2D Conv) to extract primary features from the input data. Each patch long with ReLU activation function enabling the extraction of joint spectral-spatial features. Within two 3D convolution layers, one convolution layer is used to preserve the spectral features. Additionally, the remaining 2D convolution layers are used for spatial features.

Spectral Spatial wave generation block

Figure 1 shows the spectral spatial wave generation process. This block generates the waveform representation for the given HSI patch $p \in \mathbb{R}^{S \times S \times B}$ *S*×*S* represents the patch size and *B* denotes the number of channels. Hyperspectral image data (HSI) is initially partitioned into many patches, commonly known as tokens. The characteristics of these tokens are subsequently recorded using two essential elements: the token fully connected (TFC) and the channel fully connected (CFC). Let the intermediate feature containing *n* tokens as vector$X = [x_1, x_2, x_3, \ldots, x_n]$each token$x_j$represent a vector with d-dimension. The channel-FC mathematically expressed as:

$$\mathrm{CFC}\,(x_j, W^c) = W^c x_j, j = 1, 2, 3, \ldots, n \tag{1}$$

WhereW^crepresent the learnable weights. To acquire the characteristics of each token, the channel-FC conducts a distinct process on its several channel fully connected layers, which are usually arranged in a stack with a non-linear activation function. This results in the creation of a channel-mixing MLP, which enhances the ability to transform data. In order to merge data from different tokens, the token fully connected method is necessary and outlined by:

$$\mathrm{TFC}\,(X, W^m)_j = \sum_k W^m_{jk} \odot x_k, j = 1, 2, 3, \ldots, n \tag{2}$$

WhereW^mrepresent the token mixing, $\odot$represent the element wise multiplication, and the output token *jth* determined by index *j*. The token fully connected seeks to obtain spatial information by integrating features from several tokens. By employing a basic token-mixing process with preset weights, we are constraining the capabilities of MLPs by neglecting the significant semantic information of tokens obtained from various input images. It is possible to consider each token as a wave action with two fundamental components: amplitude and phase. The primary characteristic is regarded as the magnitude, whereas the phase is computed as a multifaceted value that fluctuates depending on the semantic content of the input images. The overall result of these wave-like tokens is affected by the phase difference among them, particularly tokens that possess similar phases frequently exhibit mutual enhancement. A token represented as$\tilde{x}_j$can be seen as a wave and can be precisely defined as:

$$\tilde{x}_j = |x_j| \odot e^{i\theta j}, j = 1, 2, 3, \ldots, n \tag{3}$$

Where the imaginary part represent as $i(i^2 = -1)$. the real value defined as$|x_j|$, while$e^{i\theta j}$denotes a periodic action. The phase is denoted as$\theta_j$which represent the location information of each token within a wave. Therefore, it is regarded as a complex-valued entity for each token$\tilde{x}_j$,which consists of amplitude and phase.

When examining two tokens$\tilde{x}_1$and$\tilde{x}_2$, the amplitude(x_r) and phase(θ_r)can be precisely specified as:

$$|x_r| = \sqrt{|x_i|^2 + |x_j|^2 + 2\,|x_i| \odot |x_j| \odot \cos(\theta_j - \theta_i)} \tag{4}$$

$$x_r = \theta_i + a\tan 2(|x_j| \odot \sin(\theta_j - \theta_i), |x_i| + |x_j| \odot \cos(\theta_j - \theta_i)) \tag{5}$$

The function *atan2*(x, y) represents a tangent operation between two variables. Therefore, the amplitude of the combined result(x_r)is significantly influenced by the phase difference($\theta_j - \theta_i$).

Assume that the output feature map of the feature extractor$\hat{A} = [\hat{a}_1, \hat{a}_2, \hat{a}_3, \ldots, \hat{a}_n]$as input of the wave generation block, and the amplitude of$x_j$is defined by:

$$x_j = \mathrm{CFC}(\hat{a}_j, W^c), j = 1, 2, 3, \ldots, n \tag{6}$$

A token can be expressed as a wave that consists of real values (amplitude) and complex values (phase). token-mixing method can be used to aggregate the resultant complicated value output tokens.

$$\tilde{o}_{\mathrm{j}} = \mathrm{TFC}\left(\tilde{T}, W^m\right)_j, j = 1, 2, 3, \ldots, n \tag{7}$$

The estimation of the actual value$o_j$can be achieved by summing the real and imaginary components of the$\tilde{o}_j$, as stated by:

$$o_j = \sum_k w^m_{jk} x_k \odot \cos(\theta_k) + W^i_{jk} x_k \odot \sin(\theta_k), j = 1, 2, 3, \ldots, n, \tag{8}$$

WhereW^m, W^irepresent the learnable weights, and phase represent asθ_k.

Multilevel spectral spatial frequency generation

To fully leverage the tokens generated by the token generation module, we introduce a spectral-spatial frequency generation module that extracts high, middle, and low-frequency features, addressing the limitations of traditional Transformers in capturing fine-grained local information. While Transformers excel at modeling long-range dependencies, they tend to amplify low-frequency representations and suppress high-frequency components, which are crucial for discriminative feature extraction in HSI. To overcome this, the input tokens are first divided along the channel dimension and processed through three specialized branches. The high-frequency branch employs max pooling to preserve edge and boundary details, followed by a linear projection for adaptive feature weighting. The mid-frequency branch uses depthwise convolution to model transitional spectral features while retaining regional spatial context. The low-frequency branch performs a three-stage refinement: average pooling extracts stable, noise-suppressed base features; multi-head self-attention captures global contextual relationships; and upsampling restores spatial resolution, supported by skip connections to maintain phase alignment with other branches. This multiscale frequency-aware design allows the model to capture and fuse local and global patterns effectively, enhancing its ability to represent complex spectral-spatial variations in HSI data. From a technical perspective, the input feature maps can be represented as: $x \in \mathbb{R}^{S \times S \times B}$ and *x* decomposed into high frequency $x_h \in \mathbb{R}^{S \times S \times B_h}$, middle frequency $x_m \in \mathbb{R}^{S \times S \times B_m}$ and low frequency feature map $x_l \in \mathbb{R}^{S \times S \times B_l}$along the channel dimension, where $B = B_h + B_m + B_l$, where $B_h = B * r$and *r* denotes the channel ratio.

Since the maximum filter is highly sensitive to prominent features and the convolution operation is equally adept at capturing detailed information, we used max-pooling and linear layer operation to extract high frequency features.

$$y_h = \mathrm{FC}\left(\mathrm{MaxPool}\left(x_h\right)\right) \tag{9}$$

And for the middle frequency features linear and depthwise convolution layer is applied.

$$y_m = \mathrm{DwConv}\left(\mathrm{FC}\left(x_m\right)\right) \tag{10}$$

We use Multi-Head Self-Attention (MHSA) in the low-frequency mixer to facilitate information exchange among all tokens, leveraging its exceptional ability to extract low-frequency features. While the self-attention mechanism excels at capturing global representations, it can be computationally expensive when applied to high-resolution feature maps at shallow levels. To address this, the spatial scale of x_l is reduced by employing an average pooling layer before the self-attention operation and an upsampling layer afterward. These operations effectively lower computational costs while ensuring that the attention mechanism remains focused on embedding global information.

$$y_l = \mathrm{Upsample}\left(\mathrm{MHSA}\left(\mathrm{AvgPool}\left(x_l\right)\right)\right) \tag{11}$$

Finally, we adopted a feature fusion module to fuse the high, low and mid-level frequency features. Figure 2 shows the frequency feature fusion module followed by 3 × 3 convolutional layer.

Spectral spatial interaction module

After integrating phase, amplitude, and frequency domain features, the Spectral-Spatial Interaction Module (SSIM) is employed to enhance spectral-spatial representation through a structurally efficient and functionally rich design. SSIM begins with a depth-wise convolution, which processes each spectral channel independently, preserving spectral semantics while capturing fine-grained spatial correlations essential for local sensitivity. This is followed by two 1 × 1 convolutions: the first enables spectral interaction by projecting and mixing spectral channels, and the second refines the joint representation. in the phase, spatial interaction models smooth transitions and structural coherence between spatial tokens, while spectral evolution maintains phase continuity

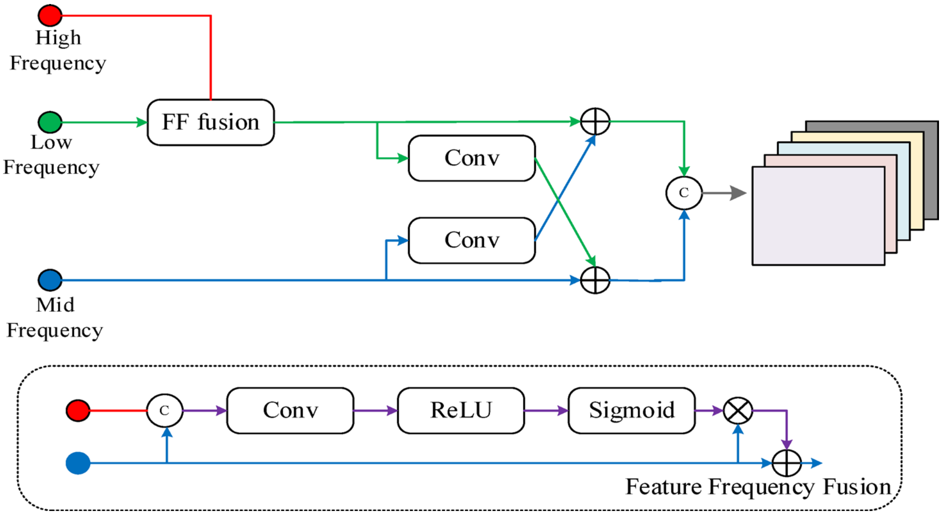


Fig. 2. Frequency Feature Fusion Module.

across bands; in the amplitude, spectral correlation reinforces similarity across bands, and spatial correlation highlights salient regions based on amplitude variations; in the frequency, high-frequency components are emphasized via inductive biases to capture less-correlated, fine-grained details, while low-frequency context provides global support, ensuring unified and comprehensive spectral-spatial modeling. This design allows SSIM to achieve effective feature fusion without relying on computationally expensive attention mechanisms. As illustrated in Fig. 1, the Spectral-Spatial Interaction module consists of two main components. The first component employs a depth-wise $\mathrm{conv}\ 3 \times 3$ layer with $\mathrm{DW}\ (.)$with the number of filters equal to the input channels. This design facilitates interactions among frequency tokens within the spatial neighborhood, effectively capturing local spatial relationships. The second component utilizes two $\mathrm{conv}\ 1 \times 1$layers$\mathrm{conv}\ (.)$, which enable individual frequencies to evolve across different spectral groups, thereby modeling spectral dependencies. This comprehensive interaction mechanism ensures that the SSIM not only preserves the fidelity of spectral-spatial features but also enhances their representation by leveraging the unique properties of the phase, amplitude, and frequency domains.

Local global modulator module

To optimize the extraction of both local and global information within the Transformer block, we introduce a Local-Global Modulator (LGM). This novel design integrates two parallel multi-scale depth-wise convolution paths and a parallel global average pooling path, as illustrated in Fig. 1. LGM processes the input feature map y_c by dividing it into two parts: one part is directed to the depth-wise convolution paths, while the other part is processed through the global average pooling path. After applying layer normalization (LN), the first half of the channels are passed through two parallel branches, utilizing 3×3 and 5×5 depth-wise convolutions to enhance the extraction of multi-scale local information. Meanwhile, the second half of the channels are directed to the global average pooling block to capture global contextual information. For an input feature map, the LGM computation is formulated as follows:

$$\tilde{y}_c = \mathrm{LN}\left(y_c\right), \left[\tilde{y}_c^1, \tilde{y}_c^2\right] = \mathrm{split}\left(\tilde{y}_c\right) \tag{12}$$

$$x_g = \mathrm{GELU}\left(\mathrm{GAP}\left(\tilde{y}_c^2\right)\right) * \tilde{y}_c^2 \tag{13}$$

$$x_l = \mathrm{GELU}\left(f_{dwc}\left(f_{1\times 1}\left(\tilde{y}_c^1\right)\right)\right) * f_{dwc}\left(f_{1\times 1}\left(\tilde{y}_c^1\right)\right) \tag{14}$$

$$y_{c+1} = y_c + \mathrm{concat}\left(x_g, x_l\right) \tag{15}$$

Dataset and experimental evaluation

To verify the performance of proposed methodology five HSI classical dataset were selected for experiments, including the Xuzhou dataset, Indian Pines, Salinas, ZY1-02D Huanghekou (ZYHHK) dataset, GF-5 Yancheng (GFYC).

Characteristics of the datasets and information summarized in the Table 1. All the dataset preprocessed following the convention to remove noisy and water absorption bands. The application of the Salinas dataset in agriculture monitoring. The number of Training and testing samples for the experiments further shown in Table 2. The application of PU dataset in agriculture, urban planning, environment monitoring and material identification due to different classes. The mineral classification is conducted on Xuzhou dataset.

Dataset	Image size	Band	class	Spatial Resolution	Spectral Resolution
IP	145 × 145	200	16	20 m	0.4–2.5 µm
SA	512 × 217	224	16	3.7 m	0.4–2.5 µm
ZYHHK	1050 × 1219	108	8	30 m	0.4–2.5 µm
GF-5	1175 × 585	147	7	30 m	0.4–2.5 µm
XU	500 × 260	436	9	0.73 m	415–2508 nm

Table 1. Dataset names and related information.

	Training Samples					Testing Samples				
Class	GF-5	ZHHK	XU	IP	SA	GF-5	ZHHK	XU	IP	SA
C1	10	31	263	5	20	350	1026	26,133	41	1989
C2	6	15	40	143	37	211	497	3987	1285	3689
C3	3	7	27	83	20	129	253	2756	747	1956
C4	24	12	52	24	14	808	400	5162	213	1380
C5	7	12	131	48	26	227	406	13,053	435	2652
C6	71	28	25	73	40	2324	913	2511	657	3919
C7	39	307	70	3	36	1266	9932	7060	25	3543
C8	–	7	48	48	112	–	241	4729	430	11,159
C9	–	–	30	2	62	–	–	3040	18	6141
C10	–	–		97	32	–	–	–	875	3246
C11	–	–		246	11	–	–	–	2209	1057
C12	–	–		59	19	–	–	–	534	1908
C13	–	–		21	9	–	–	–	184	907
C14	–	–		127	10	–	–	–	1138	1060
C15	–	–		39	72	–	–	–	347	7196
C16	–	–		9	18	–	–	–	84	1789

Table 2. Number of training and testing samples across all dataset.

Experimental setup

For this experiment, we utilized the Adam optimizer and selected categorical cross entropy as the loss function to train the suggested model. Assign a learning rate of 0.001 and weight decay of 0.0001. The batch size and epoch for experiments are set to 64 and 100, respectively. We used principle component analysis[64] for the dimensionality reduction method we select 30 spectral band after applying PCA. All experiments in this section implemented on NVIDIA GeForce RTX 3060 GPU and 64 RAM with python language framework. In order to better evaluate the classification result we choose three commonly used matrices overall accuracy (OA), average accuracy (AA) and Kappa coefficient (K).

Model parameter selection

This section examines the factors that influence classification accuracy, such as the varying patch size and PCA component.

Impact of patch size

In this experimental study, cubes of varying sizes are employed in the area of the central pixel input to maximize the utilization of spatial-spectral information in the hyperspectral Image. Classification accuracy can be influenced by the input cubes. Therefore, we conducted experiments using various patch sizes and defined the input patch size for four datasets. The overall accuracy (OA) of the proposed model increases consistently as the patch size increase for five datasets. The primary reason is the variation of the pixel distribution among the five datasets. For example, the ZYHHK dataset shows scattered land cover areas, resulting in a decrease in overall accuracy (OA) for patch size larger than 11 × 11. The smaller patch size capture local details well but may miss broader spatial context, leading to lower classification accuracy because the model does not have enough information about the surrounding area. The large patch size include more surrounding pixels, which helps the model understand the context better, increasing accuracy. Figure 3. Shows the different input patch size on five dataset.

Impact of principle component analysis

Prior to utilizing the feature extractor module, principle component analysis is applied to reduce the model parameters by decreasing the dimension of the data. Through experiments, it becomes evident that the number of principle components plays a significant role in extracting spectral spatial features. We are also conducting

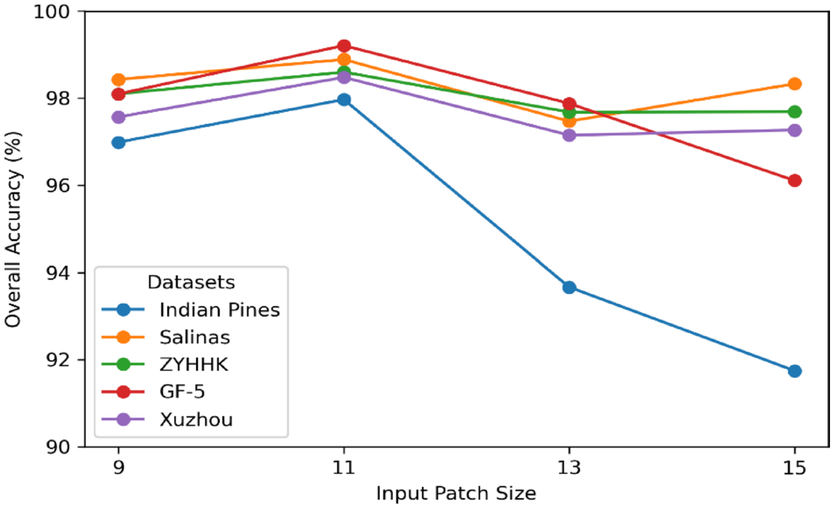


Fig. 3. OAs for five datasets with different patch sizes.

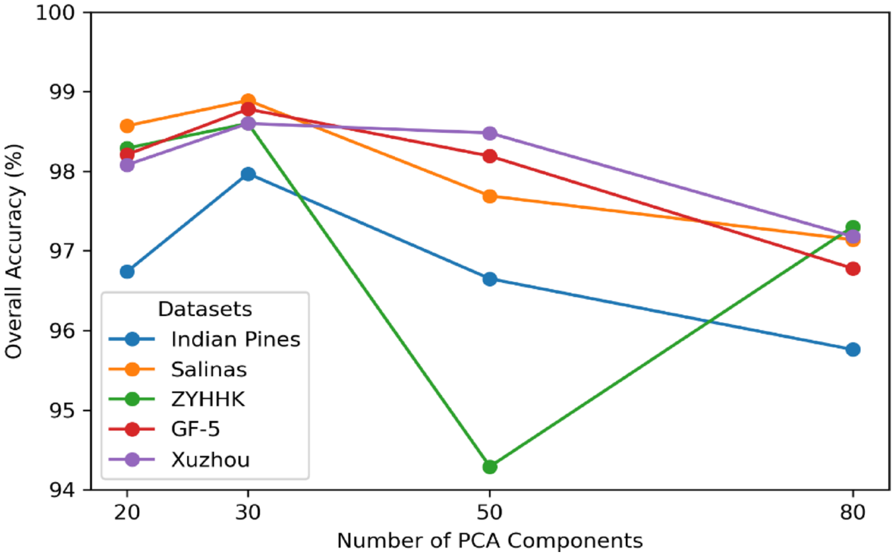


Fig. 4. OAs for five datasets with different PCA Components.

experiments to investigate the relationship between the number of spectral bands and classification accuracy. The result is present in Fig. 4.

Result and discussion

To verify the effectiveness of the proposed model, experimental results of the state-of-the art model are presented: A2S2K-Res[31], SpectralFormer[29], SSFFT[37], MorphFormer[32], GSC-ViT[33], GLMGT[34], SSMLP[35] and FTINet[36]. For a fair comparison, we use the same training samples and the same patch size for the comparison models. We conducted the experiments on f benchmark datasets, the details of which are listed in Tables 1 and 2.

Classification performance

For the Xuzhou dataset, the experimental results of each model, including OA, AA, and Kappa and class wise accuracy are shown in Table 3. Figure 5 represent the classification maps on Xuzhou dataset of each models. SpectralFormer and MorphFormer show weak performance as they fail to effectively learn Features under 1% training data, leading to clear errors in their classification maps. SSMLP and FTINet achieve high OA but exhibit low AA due to poor performance in specific classes, such as Class 6 and 7 for FTINet, which results in blocky errors. A2S2K-Res obtain > 95% in all classes which reveal that the combination of attention and CNN model extract low level features effectively. In general, SpectralFormer, MorphFormer, GLMGT, SSMLP and FTINet perform well, reducing noisy pixels in the maps; however, Our proposed model approach surpasses in category 6, 7, and 9 than other advanced techniques, achieving the highest AA, OA, and KAPPA scores 98.48%, 98.75%

Class	A2S2K-Res[31]	SF[29]	SSFTT[37]	MorphForm[32]	GSC-ViT[33]	GLMGT[34]	SSMLP[35]	FTINet[36]	Proposed
C1	97.55(± 1.04)	96.23(± 0.03)	97.99(± 0.07)	97.82(± 0.09)	98.04(± 0.32)	97.35(± 0.06)	98.36(± 0.02)	**98.43**(± 0.06)	98.2(± 0.18)
C2	97.91(± 1.10)	97.61(± 0.6)	98.47(± 0.02)	97.79(± 0.31)	**99.59(± 0.16)**	99.34(± 0.03)	98.54(± 0.62)	99.09(± 0.03)	99.11(± 0.02)
C3	98.03(± 0.93)	92.41(± 0.3)	96.18(± 1.08)	94.41(± 0.05)	96.58(± 1.26)	89.98(± 5.01)	98.91(± 0.32)	98.98(± 1.02)	97.61(± 0.04)
C4	98.14(± 1.05)	94.69(± 0.7)	98.29(± 0.03)	97.84(± 0.22)	97.22(± 1.02)	**99.65(± 0.05)**	96.02(± 0.01)	95.37(± 0.05)	97.92(± 0.01)
C5	99.29(± 0.01)	99.34(± 0.08)	99.03(± 0.07)	98.88(± 0.13)	98.62(± 0.02)	**99.77(± 0.03)**	98.94(± 0.11)	99.13(± 0.03)	97.88(± 0.06)
C6	97.13(± 1.36)	92.08(± 0.5)	99.21(± 0.09)	98.09(± 0.09)	98.42(± 0.35)	97.92(± 0.37)	99.00(± 0.02)	98.75(± 0.37)	**99.23(± 0.05)**
C7	98.33(± 0.83)	98.54(± 0.4)	99.13(± 0.05)	95.75(± 0.04)	99.07(± 0.57)	99.29(± 0.05)	98.97(± 0.16)	99.14(± 0.05)	**99.35(± 0.04)**
C8	99.49(± 0.06)	97.56(± 0.3)	98.16(± 0.02)	99.32(± 0.16)	97.35(± 1.25)	**100(± 0.00)**	99.06(± 0.01)	99.72(± 0.01)	99.95(± 0.01)
C9	97.13(± 1.03)	95.95(± 0.25)	97.69(± 0.08)	90.22(± 4.18)	96.05(± 0.37)	96.57(± 0.20)	97.07(± 0.07)	97.07(± 0.20)	**99.55(± 0.04)**
OA	98.15(± 0.17)	96.80(± 0.32)	98.34(± 0.23)	97.45(± 0.25)	98.10(± 0.59)	98.19(± 0.13)	98.40(± 0.05)	98.31(± 0.03)	**98.48(± 0.18)**
AA	98.11(± 0.03)	96.04(± 0.74)	98.29(± 0.30)	96.68(± 0.11)	97.88(± 0.46)	97.76(± 0.02)	98.31(± 0.20)	98.40(± 0.02)	**98.75(± 0.25)**
K×100	97.65(± 0.08)	95.95(± 0.85)	97.90(± 0.42)	96.76(± 0.16)	97.60(± 0.20)	97.70(± 0.11)	97.98(± 0.15)	98.11(± 0.05)	**98.08(± 0.31)**
Param(K)	105.4 K	120.92 K	153.8 K	141.1 K	83.02 K	104.44 K	291.01 K	260.11 K	514.28 K
MACs(M)	35.22 M	3.42 M	6.99 M	5.97 M	9.43 M	12.39 M	34.06 M	31.16 M	2.45 M
TR Time(s)	71.98	39.78	36.50	101.15	116.94	91.38	72.08	90.02	162.32
TS Time(s)	8.33	37.00	10.05	36.10	22.38	9.26	7.50	9.2	15.87

Table 3. Classification result (%) on Xuzhou Dataset.

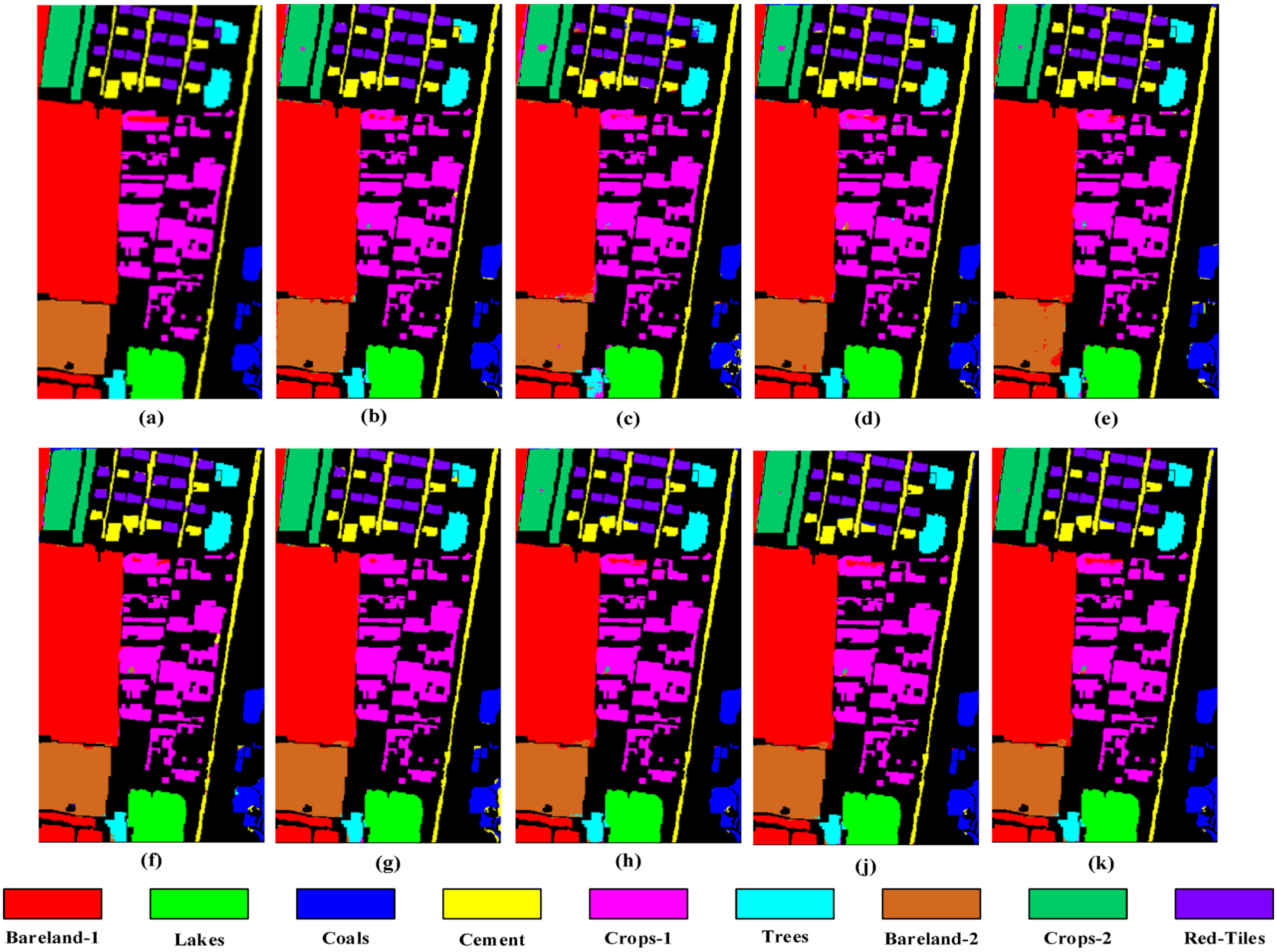


Fig. 5. Classification Maps obtained by different Models on Xuzhou Dataset. (a) Ground Truth. (b) A2S2K-Res. (c) SpectralFormer. (d) SSFTT. (e) MorphFormer. (f) GSC_ViT. (g) GLMGT. (h) SSMLP. (j) FTINet. (k) Proposed.

Class	A2S2K-Res[31]	SF[29]	SSFTT[37]	MorphFormer[32]	GSC-ViT[33]	GLMGT[34]	SSMLP[35]	FTINet[36]	Proposed
C1	92.68(± 1.80)	96.58(± 3.14)	96.58(± 2.30)	95.41(± 3.11)	97.07(± 0.46)	**98.92**(± 1.01)	97.17(± 1.02)	96.87(± 3.90)	88.1(± 3.90)
C2	83.90(± 5.04)	72.63(± 3.91)	89.93(± 10.02)	82.69(± 10.89)	98.99(± 0.77)	94.76(± 1.60)	91.34(± 5.10)	98.18(± 0.39)	**100**(± 0.00)
C3	82.93(± 4.02)	69.84(± 12.25)	81.74(± 8.05)	76.58(± 11.07)	82.93(± 10.4)	74.60(± 12.63)	81.34(± 6.63)	87.30(± 6.10)	**95.58**(± 4.12)
C4	100(± 0.00)	95.25(± 3.06)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)
C5	72.59(± 12.2)	62.22(± 7.50)	72.09(± 14.0)	75.80(± 9.05)	71.35(± 8.94)	55.06(± 13.80)	69.62(± 9.30)	68.64(± 3.01)	**93.98**(± 4.92)
C6	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)	100(± 0.00)
C7	99.90(± 0.10)	100(± 0.00)	99.97(± 0.03)	99.53(± 0.47)	100(± 0.00)	99.81(± 0.19)	100(± 0.00)	100(± 0.00)	99.82(± 0.18)
C8	81.32(± 11.2)	52.69(± 13.60)	87.55(± 10.08)	80.49(± 5.60)	89.21(± 9.09)	87.96(± 9.42)	83.40(± 8.03)	84.23(± 1.19)	**92.98**(± 0.11)
OA	97.34(± 0.60)	96.09(± 3.61)	97.98(± 0.26)	97.19(± 1.32)	98.39(± 0.40)	97.58(± 2.11)	97.93(± 0.7)	98.25(± 5.36)	**98.6**(± 1.80)
AA	89.16(± 1.30)	81.15(± 11.10)	90.98(± 5.84)	88.81(± 3.81)	92.44(± 5.94)	88.89(± 8.74)	90.36(± 2.65)	91.90(± 8.03)	**96.3**(± 3.30)
K×100	94.18(± 2.90)	91.17(± 5.06)	95.57(± 3.89)	93.85(± 0.56)	96.46(± 3.52)	94.72(± 1.46)	95.47(± 3.76)	96.17(± 3.75)	**96.94**(± 0.05)
Param(K)	105.37 K	120.85 K	153.7 K	141.03 K	82.89 K	104.4 K	291.72 K	260.11 K	498.37 K
MACs(M)	35.22 M	3.42 M	6.99 M	5.97 M	9.43 M	12.39 M	34.06 M	31.16 M	1.39 M
TR Time(s)	93.26	42.06	47.73	134.70	153.21	64.37	53.86	59.74	174.36
TS Time(s)	3.47	5.09	2.37	5.92	6.45	4.88	2.42	2.36	10.51

Table 4. Classification result (%) on ZYHHK dataset.

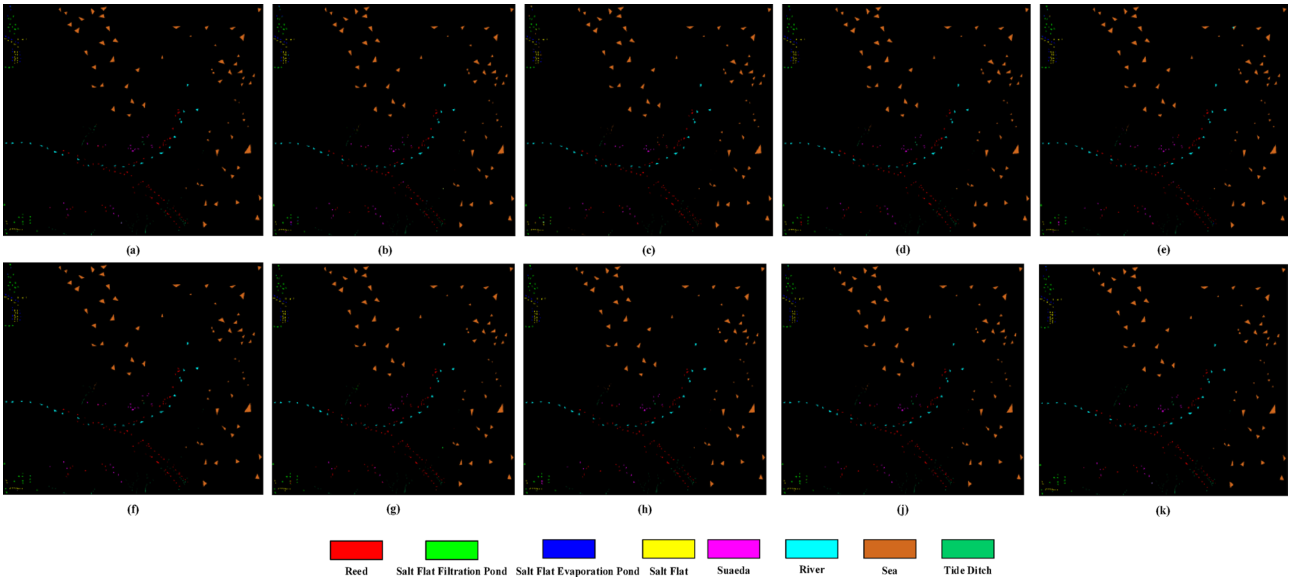


Fig. 6. Classification Maps obtained by different Models on ZYHHK Dataset. (a) Ground Truth. (b) A2S2K-Res. (c) SpectralFormer. (d) SSFTT. (e) MorphFormer. (f) GSC_ViT. (g) GLMGT. (h) SSMLP. (j) FTINet. (k) Proposed.

and 98.08%, with all class accuracies exceeding 98%. Additionally, the classification map generated by our method is the most precise and smooth, closely resembling the ground-truth image.

Table 4 shows the classification results for the ZYHHK dataset. The results revealed that the proposed model exhibits constant performance gain in all the classes. It can observed that the OA achieved by the SpectralFormer, MorphFormer and GLMGT model is slightly low. Focusing on the GSC-ViT and FTINet models outperformed and gained the highest overall accuracy. The proposed model obtained better classification results in class 2, 3, 5 and 8 where the both wave generation and multilevel frequency generation layer helps to capture frequency features better. Compared to the second-best model, the proposed model achieved 0.35%, 4.4%, and 0.77% in terms of overall accuracy, average accuracy, and kappa coefficient. Figure 6. Shows the classification maps of different methods on the ZYHHK Dataset; as we can see, the FTINet model and proposed model have less noise. Our classification is almost near to the ground truth image. As the level of noise increases, the accuracy of the classification maps tends to decrease.

Table 5 shows the result of all metrics on Salinas dataset and Fig. 7 shows the classification maps obtained by all models on Salinas dataset. Our model demonstrates outstanding classification performance, achieving AA, OA, and KAPPA scores exceeding 99.07%, 98.93%, and 0.98, respectively—the highest among all compared methods. Specifically, the MorphFormer, SpectralFormer and GLMGT models continue to underperform due

Class	A2S2K-Res[31]	SF[29]	SSFTT[37]	MorphFormer[32]	GSC-ViT[33]	GLMGT[34]	SSMLP[35]	FTINet[36]	Proposed
C1	100(±0.00)	98.79(±0.90)	100(±0.00)	99.94(±0.06)	100(±0.00)	98.03(±0.84)	100(±0.00)	99.94(±0.06)	99.32(±0.06)
C2	100(±0.00)	100(±0.00)	99.91(±0.08)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	99.97(±0.03)
C3	100(±0.00)	99.59(±0.06)	99.79(±0.21)	99.84(±0.16)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	99.84(±0.16)
C4	**99.92**(±0.08)	95.0(±3.03)	99.63(±0.14)	96.66(±0.16)	99.56(±0.44)	75.14(±6.99)	99.56(±0.44)	99.63(±0.02)	93.46(±1.46)
C5	98.22(±0.06)	96.30(±1.40)	96.60(±2.57)	99.32(±0.10)	96.26(±0.33)	100(±0.00)	99.96(±0.04)	99.54(±0.06)	99.65(±0.35)
C6	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)
C7	100(±0.00)	99.88(±0.06)	99.97(±0.03)	99.94(±0.06)	99.88(±0.22)	100(±0.00)	100(±0.00)	99.94(±0.06)	99.75(±0.25)
C8	99.50(±0.12)	92.15(±4.36)	98.17(±0.44)	97.52(±1.12)	98.37(±0.02)	97.19(±1.10)	94.44(±1.00)	93.81(±0.35)	**99.88**(±0.12)
C9	100(±0.00)	100(±0.00)	100(±0.00)	99.90(±0.10)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)
C10	99.19(±0.26)	97.04(±1.25)	99.38(±0.04)	99.32(±0.68)	**99.50**(±0.50)	99.16(±0.55)	99.16(±0.84)	99.22(±0.28)	96.67(±1.28)
C11	99.71(±0.05)	93.94(±0.27)	99.90(±0.10)	91.20(±5.19)	99.43(±0.57)	99.90(±0.10)	99.90(±0.10)	99.81(±0.19)	100(±0.00)
C12	100(±0.00)	99.63(±0.20)	100(±0.00)	99.10(±0.90)	99.84(±0.10)	100(±0.00)	100(±0.00)	100(±0.00)	99.27(±0.73)
C13	82.02(±0.28)	97.24(±0.11)	94.92(±1.65)	85.44(±6.58)	82.24(±5.89)	3.30(±9.81)	95.03(±2.60)	94.92(±1.03)	**99.88**(±0.12)
C14	98.01(±1.08)	96.22(±1.81)	98.86(±1.26)	93.86(±2.09)	99.62(±0.38)	77.14(±8.73)	98.58(±1.30)	98.39(±1.79)	98.39(±0.13)
C15	93.71(±2.03)	84.54(±4.84)	93.41(±5.27)	93.03(±4.34)	99.58(±0.17)	91.95(±6.66)	94.94(±3.50)	95.70(±4.38)	**96.74**(±0.56)
C16	99.44(±0.09)	98.65(±1.13)	99.44(±0.56)	99.67(±0.33)	99.70(±0.30)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)
OA	98.54(±0.18)	95.43(±1.56)	98.37(±0.79)	97.78(±1.00)	99.02(±0.06)	95.48(±0.34)	97.98(±2.01)	97.93(±1.60)	**99.07**(±0.06)
AA	98.11(±1.31)	96.81(±2.06)	98.75(±0.61)	97.08(±2.35)	98.31(±0.55)	90.11(±1.85)	98.85(±0.06)	98.81(±0.73)	**98.93**(±0.26)
K×100	98.38(±0.63)	94.91(±3.16)	98.19(±0.13)	97.53(±3.15)	98.93(±0.41)	94.96(±3.05)	97.75(±0.08)	97.69(±1.90)	**98.96**(±0.34)
Param(K)	105.57 K	121.7 K	154.26 K	141.55 K	83.25 K	104.67 K	291.72 K	189.45 K	505.75 K
MACs(M)	35.22 M	3.42 M	6.99 M	5.98 M	9.43 M	12.39 M	34.06 M	12.58 M	1.35 M
TR Time(s)	116.13	35.19	58.51	174.54	178.56	38.55	72.26	74.61	258.61
TS Time(s)	8.33	13.45	5.40	19.30	19.68	9.84	5.74	5.35	32.45

Table 5. Classification result (%) on Salinas dataset.

to using only 1% of labeled samples for training. The basic reason for this is because spectralFormer model do not fully exploit the three- dimensional nature of the data, and it is challenging to implicitly represent the spatial relationship with a limited number of training samples. Out of many comparative algorithms, GSC-ViT, SSMLP, GLMGT and FTINet are particularly well suited for hyperspectral image data and shows the superior performance since they employ convolutional layers. Hybrid CNN and transformer approaches outperform transformer approaches. This implies that the combination of two architectures can be advantageous, but it requires more refinement. Furthermore, the proposed model effectively extracts spectral spatial frequency features specifically designed for hyperspectral image data. It successfully extracts both local and high-level frequency features by combining the strength of the convolutional neural network and transformer-based frequency domain learning module. As a result, it outperforms other transformer-based classification models, providing a significant advantage. Comparison models may exhibit poor performance while hyperspectral image classification using limited number of samples.

The classification results of GF-5 dataset are shown in Table 6. Hybrid CNN and transformer approaches outperform transformer approaches. This implies that the combination of two architectures can be advantageous to extract features, but it requires more refinement. It successfully extracts both local and high-level features by combining the strength of the convolutional neural network and transformer-based attention module. As a result, it outperforms other transformer-based classification models, providing a significant advantage. Comparison models may exhibit poor performance. The proposed model achieves the best classification result in all classes specifically in class 3 and 5 achieved best classification results. Figure 8. Shows the classification maps on the GF-5 dataset. As one can see, SSMLP and FTINet obtain excellent classification results with less noise and intra-class smoothness. Furthermore, it is evident that the classification performance of proposed model surpasses that of the four global-local models (i.e. SSFTT, GSC_ViT GLMGT and FTINet) across all five datasets. For instance, on the IP dataset Table 7; Fig. 9, proposed model achieves an OA value that is 0.8% higher than FTINet and 1.09% higher than SSMLP.

Model complexity analysis

Tables 3, 4, 5, 6 and 7 shows the model complexity, revealing that the proposed model is less complex in terms of FLOPs the parameters of the proposed model is higher than GLMGT, SSMLP and FTINet model. as we can see the training time of the proposed model is lower than other transformer based models except SSFTT, in terms of testing time SSMLP and A2S2K-Res model is perform well the parameters of this model are also very low. We also report the training and testing time obtained by different models on five datasets. As shown in Tables 3, 4, 5, 6 and 7 the proposed model has highest parameters than other models but also achieving significantly shorter training and testing time than some other models MorphFormer, GSC_ViT, SSMLP and FTINet model. SpectralFormer and HybridSN models demonstrated that the faster speed while maintaining performance ≥ 90%. On the GF-5 dataset the proposed model complete training in 91.24 s.

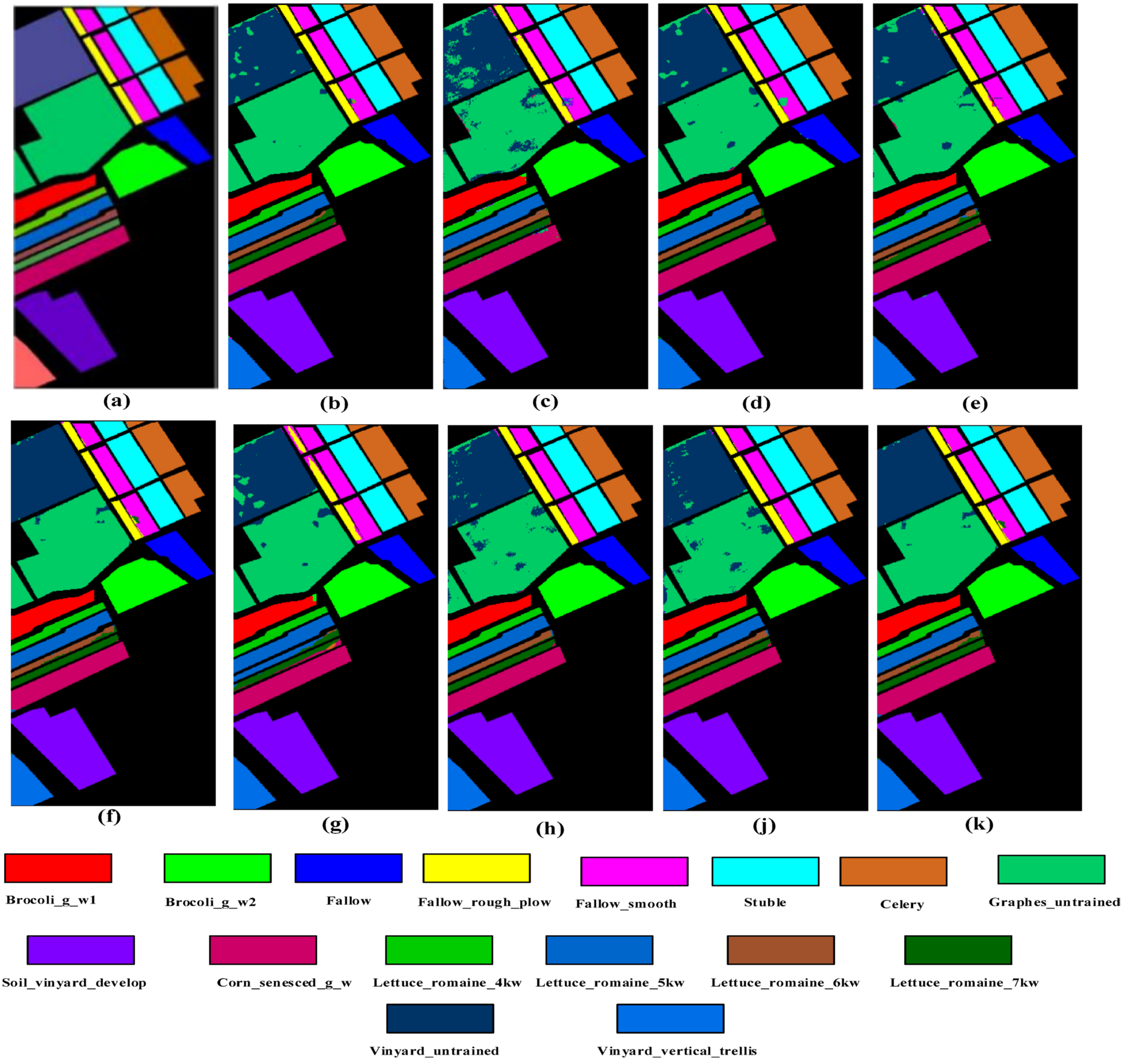


Fig. 7. Classification Maps obtained by different Models on Salinas Dataset. (a) Ground Truth. (b) A2S2K-Res. (c) SpectralFormer. (d) SSFTT. (e) MorphFormer. (f) GSC_ViT. (g) GLMGT. (h) SSMLP. (j) FTINet. (k) Proposed.

Ablation experiments

To comprehensively highlight the efficacy of each module, we conducted various combinations on the four dataset using patch size 11 × 11, and same training samples. Our analysis focuses on the wave generation, multilevel frequency generation SSIM module and LGM module, accessed through OA, AA and Kappa metrics. The outcomes of these experiments are presented in Table 8. Overall, the proposed network can achieve the best performance results. In Case 1, directly input the HSI data into transformer which utilizes wave generation, frequency generation, SSIM and LGM module, yields a relatively low Overall Accuracy (OA) value due to not using feature extraction module. In contrast, Cases 2 and 3 demonstrate significant improvements in OA accuracy by incorporating token generation module, respectively and wave generation module are not using. Notably, Case 4 and 5 further enhances performance by combining features from both branches using an additive approach. These improvement highlight the effectiveness of the wave generation and frequency generation module in enhancing the model capability to capture spectral spatial complexities of hyperspectral data in the frequency domain, resulting in better classification performance. In conclusion, the ablation experiments clearly demonstrate the substantial impact of each module in improving the model performance, especially activate in addressing spatial complexities and enhancing overall classification accuracy.

Class	A2S2K-Res[31]	SF[29]	SSFTT[37]	MorphFor[32]	GSC-ViT[33]	GLMGT[34]	SSMLP[35]	FTINet[36]	Proposed
C1	100(±0.00)	77.65(±10.55)	98.85(±1.40)	92.55(±2.3)	98.85(±0.17)	100(±0.00)	97.99(±2.01)	99.14(±0.86)	98.65(±0.03)
C2	41.70(±10.20)	45.97(±13.51)	78.19(±10.03)	45.97(±11.25)	81.04(±9.0)	52.60(±30.16)	85.78(±7.8)	77.25(±14.02)	67.8(±10.37)
C3	53.125(±15.8)	63.28(±12.36)	66.40(±20.6)	56.25(±10.82)	57.03(±13.02)	57.03(±24.32)	65.62(±19.95)	66.40(±21.31)	**100**(±0.00)
C4	99.13(±0.7)	98.88(±0.83)	99.62(±0.3)	98.88(±0.84)	97.89(±0.08)	100(±0.00)	99.87(±0.03)	99.25(±1.85)	97.78(±0.05)
C5	74.88(±8.76)	62.99(±10.76)	94.27(±0.7)	86.78(±0.95)	87.66(±0.23)	56.82(±18.16)	96.47(±1.07)	94.71(±0.06)	**100**(±0.00)
C6	99.78(±0.22)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)
C7	100(±0.00)	100(±0.00)	100(±0.00)	99.44(±0.07)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)
OA	95.25(±1.04)	93.74(±1.03)	97.94(±0.48)	95.44(±0.02)	97.28(±0.30)	95.23(±1.26)	98.10(±0.05)	97.89(±0.03)	**98.29**(±0.01)
AA	81.23(±3.10)	78.39(±1.47)	91.05(±0.5)	82.84(±1.08)	88.92(±0.42)	80.92(±1.02)	92.25(±2.04)	90.96(±1.01)	**94.89**(±0.20)
K×100	93.39(±2.93)	91.14(±1.57)	97.14(±0.4)	93.61(±0.03)	96.23(±0.09)	93.34(±0.02)	97.24(±0.03)	97.06(±0.05)	**97.61**(±0.13)
Param(K)	105.35 K	120.79 K	153.74 K	140.97 K	82.98 K	104.28 K	290.81 K	189.45 K	498.36 K
MACs(M)	35.22 M	3.42 M	6.99 M	5.97 M	9.43 M	12.39 M	34.06 M	12.58 M	1.39 M
TR Time(s)	29.46	20.26	23.16	52.98	69.06	30.68	28.69	21.02	91.24
TS Time(s)	0.92	0.64	1.20	4.60	2.61	2.07	1.21	3.21	5.21

Table 6. Classification result (%) on GF-5 dataset.

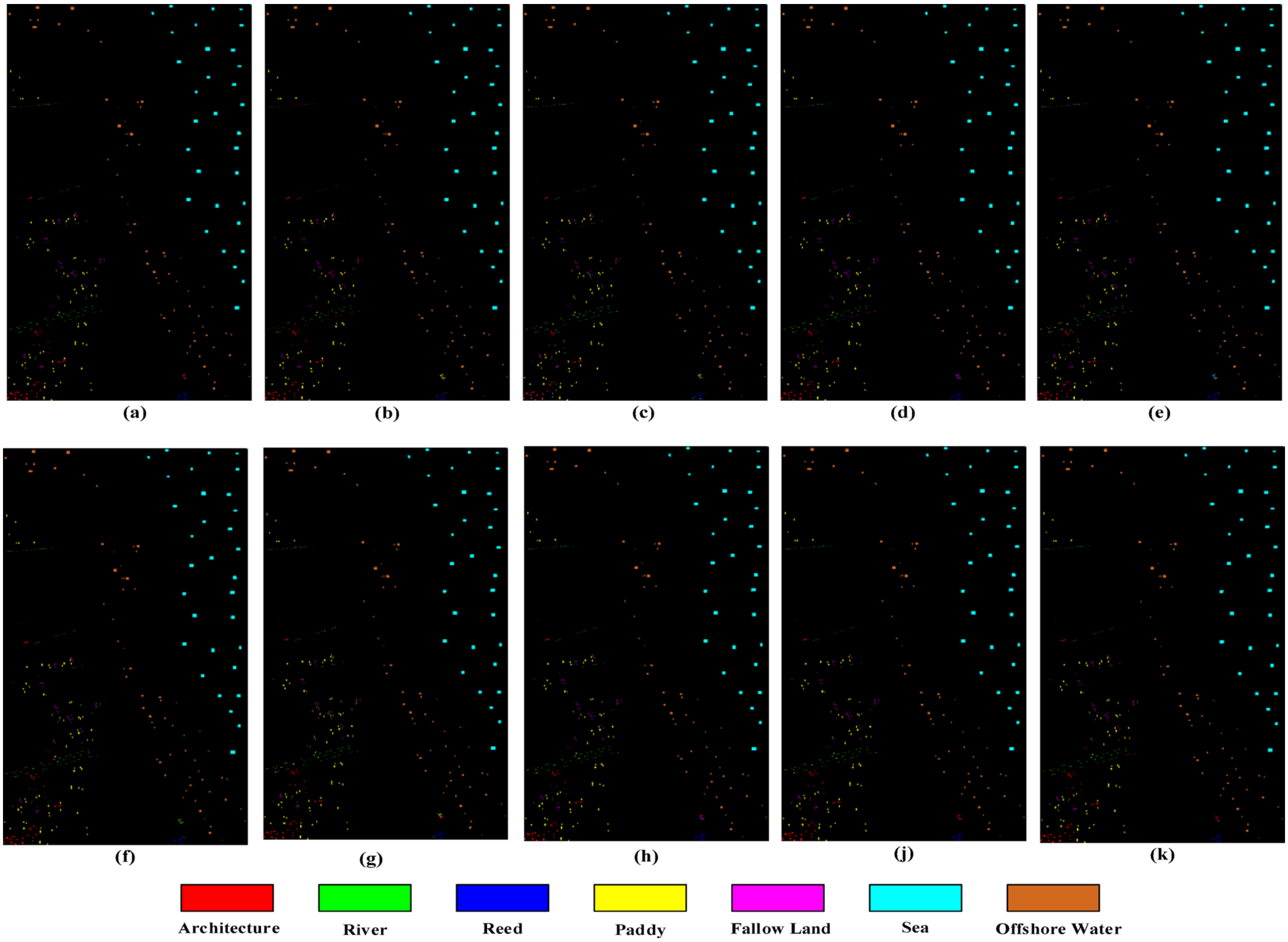


Fig. 8. Classification Maps obtained by different Models on GF-5 Dataset. (a) Ground Truth. (b) A2S2K-Res. (c) SpectralFormer. (d) SSFTT. (e) MorphFormer. (f) GSC_ViT. (g) GLMGT. (h) SSMLP. (j) FTINet. (k) Proposed.

Figure 10(a) and 10(b) illustrate that increasing PCA component values result in exponential growth in computational time and parameter requirements. Consequently, we selected 30PCA component as the optimal PCA dimensionality reduction parameter for our proposed method, striking a balance between performance and computational efficiency.

Class	A2S2K-Res[31]	SF[29]	SSFTT[37]	MorphFor[32]	GSC-ViT[33]	GLMGT[34]	SSMLP[35]	FTINet[36]	Proposed
C1	36.57(±20.02)	14.63(±32.05)	81.39(±4.82)	13.63(±5.83)	95.34(±5.99)	65.47(±8.72)	95.12(±2.31)	97.56(±1.74)	97.56(±1.07)
C2	95.79(±1.11)	76.42(±12.04)	92.62(±6.90)	89.90(±9.97)	95.18(±3.80)	94.0(±1.67)	91.43(±0.33)	92.45(±0.39)	97.78(±0.03)
C3	99.19(±0.7)	74.43(±11.04)	99.74(±0.04)	85.42(±13.16)	98.44(±0.31)	99.86(±0.14)	97.45(±1.06)	99.46(±0.56)	94.62(±2.21)
C4	96.24(±1.62)	38.02(±17.18)	82.95(±6.50)	94.22(±3.31)	95.00(±4.12)	86.85(±2.10)	96.71(±1.03)	97.18(±1.05)	94.81(±0.32)
C5	98.39(±0.32)	91.72(±6.25)	100(±0.00)	99.34(±0.66)	99.10(±0.00)	95.17(±2.44)	95.86(±0.90)	95.68(±1.10)	100(±0.00)
C6	99.54(±0.01)	98.17(±0.21)	99.41(±0.11)	99.13(±0.87)	96.90(±2.02)	99.08(±0.60)	98.78(±0.96)	99.08(±0.92)	100(±0.00)
C7	16.87(±21.11)	58.34(±14.01)	84.61(±4.80)	48.14(±21.39)	7.69(±35.49)	56.17(±11.8)	100(±0.00)	88(±12.43)	92.3(±0.36)
C8	100(±0.00)	99.06(±0.62)	100(±0.00)	99.77(±0.23)	100(±0.00)	100(±0.00)	100(±0.00)	100(±0.00)	99.29(±0.7)
C9	41.57(±10.16)	11.11(±31.04)	36.84(±19.4)	36.84(±11.9)	31.57(±12.26)	48.98(±17.8)	44.44(±20.8)	47.68(±26.0)	100(±0.00)
C10	99.77(±0.01)	79.08(±12.65)	97.15(±1.38)	93.60(±3.72)	93.91(±5.08)	98.74(±0.82)	93.25(±1.74)	96.8(±1.04)	93.66(±0.02)
C11	99.27(±0.07)	90.40(±8.09)	97.09(±2.56)	97.04(±1.59)	98.68(±0.29)	99.32(±0.68)	98.59(±1.48)	97.19(±1.30)	99.33(±0.08)
C12	98.68(±0.15)	52.05(±13.01)	91.93(±4.45)	76.19(±16.08)	91.66(±5.99)	97.94(±1.62)	97.19(±2.54)	98.12(±0.56)	97.36(±2.96)
C13	97.29(±0.40)	97.83(±1.30)	96.37(±3.51)	96.92(±2.65)	89.00(±3.50)	97.29(±2.49)	98.91(±0.01)	98.37(±0.39)	95.81(±0.87)
C14	99.64(±0.35)	96.83(±3.50)	99.91(±0.09)	99.58(±0.42)	100(±0.00)	100(±0.00)	99.91(±0.09)	99.91(±0.58)	99.55(±0.45)
C15	99.42(±0.18)	69.16(±21.40)	94.76(±3.25)	94.82(±3.20)	97.77(±2.07)	97.98(±1.30)	98.84(±1.11)	98.84(±0.47)	97.86(±1.27)
C16	83.33(±3.02)	83.33(±5.94)	89.65(±7.40)	76.13(±10.05)	90.69(±4.22)	69.04(±10.8)	98.80(±1.19)	91.66(±0.33)	98.75(±0.02)
OA	96.74(±2.04)	83.02(±1.46)	96.46(±2.91)	93.30(±2.22)	96.65(±2.02)	96.83(±2.23)	96.88(±2.04)	97.17(±1.05)	97.97(±1.14)
AA	85.09(±5.01)	70.66(±9.70)	90.27(±5.56)	81.29(±9.68)	86.30(±10.18)	87.86(±14.0)	94.08(±2.06)	93.62(±0.44)	97.41(±0.82)
K×100	96.74(±0.16)	80.43(±8.50)	95.96(±3.25)	92.34(±3.78)	96.17(±1.41)	96.38(±2.95)	96.44(±0.94)	91.66(±2.10)	97.67(±2.09)
Param(K)	105.57 K	121.7 K	154.26 K	141.55 K	83.25 K	104.67 K	291.72 K	189.45 K	504.62 K
MACs(M)	35.22 M	3.42 M	6.99 M	5.98 M	9.43 M	12.39 M	34.06 M	12.58 M	1.3 M
TR Time(s)	214.99	54.19	98.01	124.68	235.34	131.93	130.87	87.54	242.14
TS Time(s)	2.74	1.59	1.82	5.54	5.70	2.51	1.80	5.36	6.55

Table 7. Classification result (%) on IP dataset.

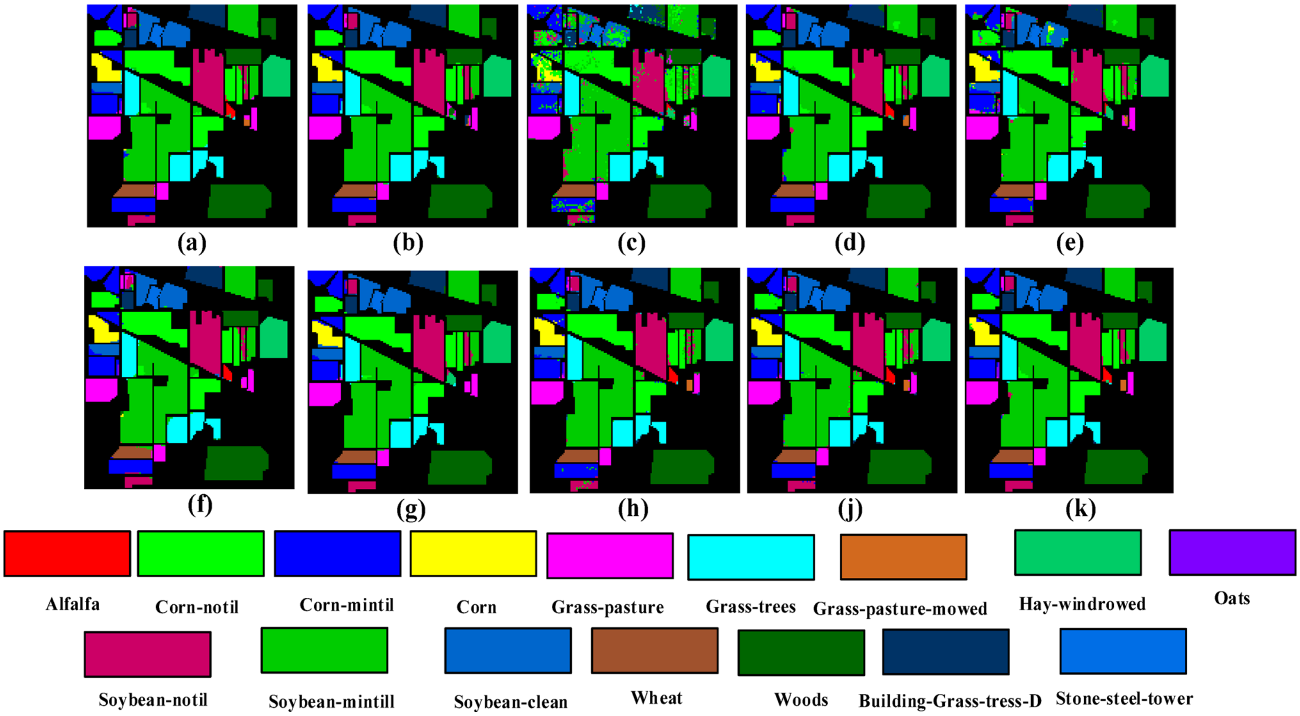
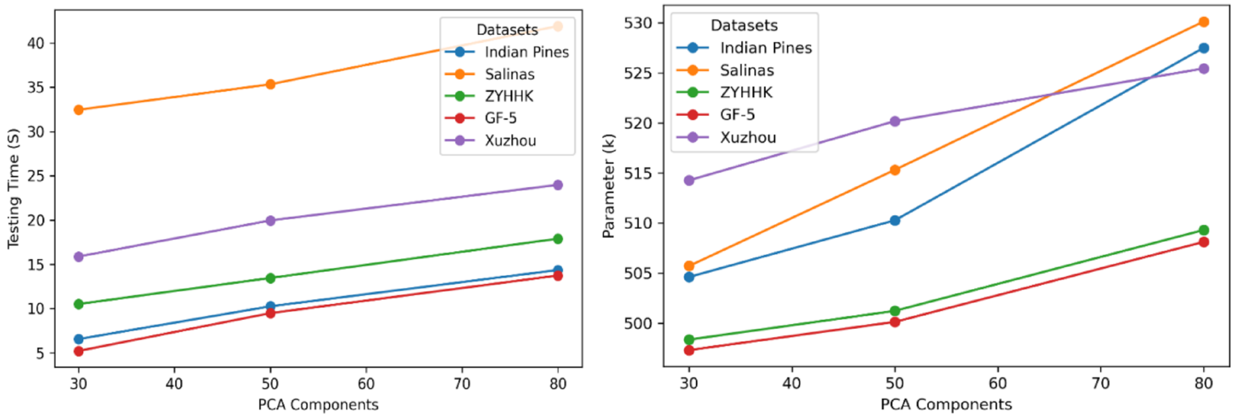


Fig. 9. Classification Maps obtained by different Models on Indian Pines Dataset. (a) Ground Truth. (b) A2S2K-Res. (c) SpectralFormer. (d) SSFTT. (e) MorphFormer. (f) GSC_ViT. (g) GLMGT. (h) SSMLP. (j) FTINet. (k) Proposed.

Component					**Dataset**				
Token Generation	**Wave Generation**	**Frequency Generation**	**SSIM**	**LGM**	**XU**	**IP**	**SA**	**GF-5**	**ZYHHK**
×	✓	✓	✓	✓	93.65	90.57	95.20	89.31	90.23
✓	×	✓	✓	✓	95.28	91.98	96.06	90.58	90.89
✓	✓	×	✓	✓	96.05	93.61	97.47	91.63	91.69
✓	✓	✓	×	✓	97.36	95.11	97.88	95.36	92.88
✓	✓	✓	✓	×	97.87	96.77	98.22	97.19	95.24
✓	✓	✓	✓	✓	**98.48**	**97.97**	**99.07**	**98.29**	**98.60**

Table 8. Ablation experiments (%) with different component on five dataset.



(a) **(b)**

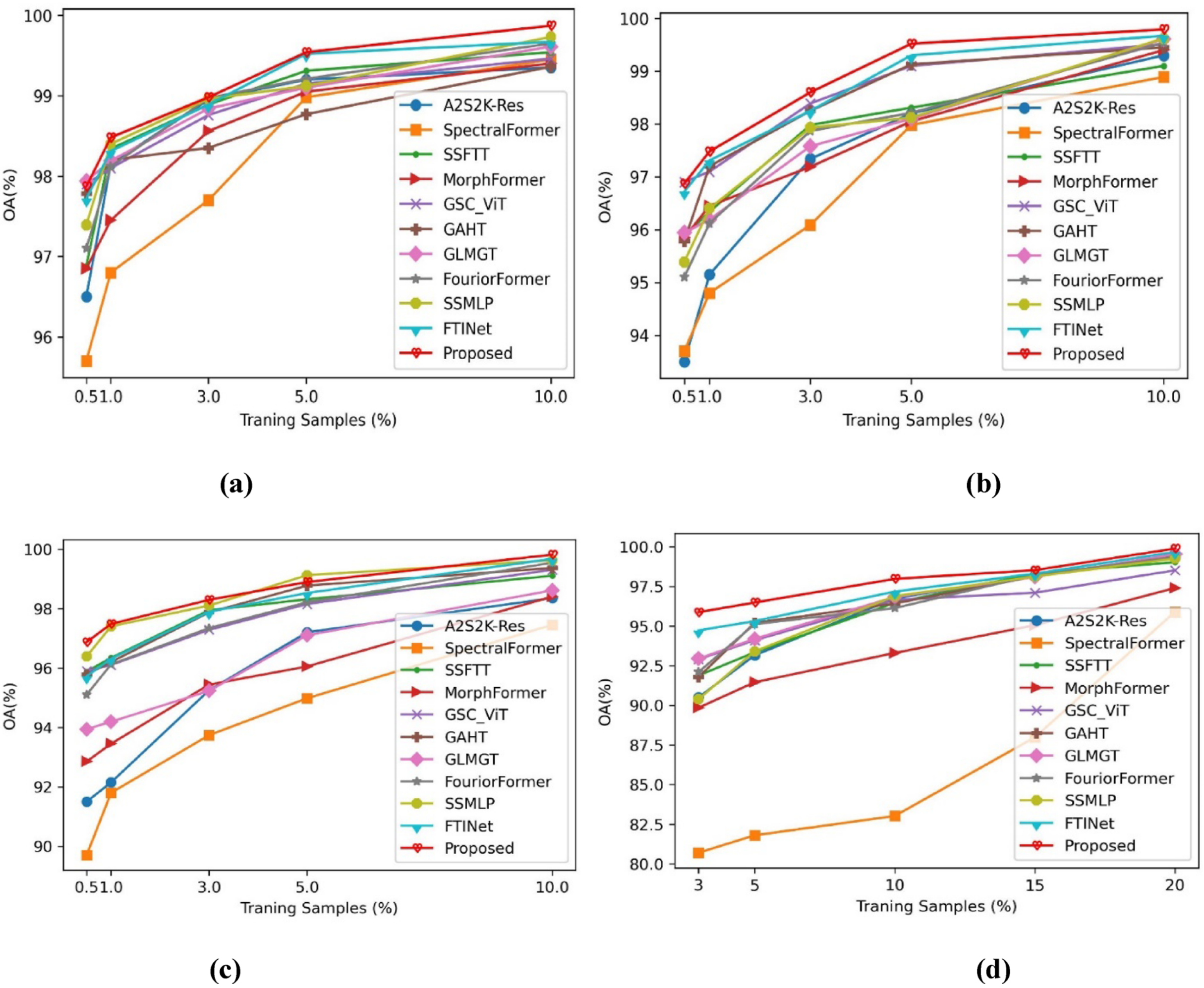
Fig. 10. Comparison of different PCA component. (a) Impact of different PCA component on testing time. (b) Impact of different PCA component on Parameters.

Impact of training samples

The proportion of training samples significantly influences hyperspectral image (HSI) classification performance. However, the scarcity of labeled samples hinders model training. To address this, we evaluated our method's effectiveness using limited training samples. We conducted experiments on the four datasets, utilizing 0.5%, 1%, 5%, and 10% of the available samples for Xuzhou, ZYHHK and GF-5 dataset for the IP dataset 3%,5%10% and 20% samples are used. The results, presented in Fig. 11, demonstrate that our method consistently achieves the highest Overall Accuracy (OA) values across various training sample percentages. Notably, our approach surpasses 95% OA on both datasets with merely 0.5% of the samples. These findings confirm that our method maintains its superiority even when faced with limited training data. Furthermore, it can be observed that proposed model stabilized faster than other models. SSMLP, FTINet, GSC_ViT, GAHT and SSFTT also stabilize but with at different OA accuracies points.

Conclusion

In this article, we proposed a novel model designed to effectively handle HSIC tasks in the frequency domain by utilizing the token generation module, a spectral spatial wave generation module, spectral spatial frequency generation module, SSIM and a LGM module. The model's primary goal is to extract deep spectral-spatial features in the frequency domain, which are critical for improving classification accuracy. The frequency transformer encoder block serves as the backbone of the model, token generation module functioning as a robust feature extractor in the low-level spectral-spatial frequency features. To capture high-level semantic features, we proposed the transformer encoder block. In the wave generation block the amplitude representation encodes the intensity of the features, reflecting the overall strength of the spectral-spatial feature, while the phase representation captures positional and structural information, highlighting spatial relationships and variations in the spectral domain. Together, these wave tokens, combining both phase and amplitude information, form the basis for subsequent multilevel frequency feature extraction. In multilevel frequency generation block High-frequency features capture fine-grained local details crucial for distinguishing small-scale variations in spectral-spatial relationships. Mid-frequency features act as a bridge between high- and low-frequency features, capturing medium-scale patterns and interactions. Low-frequency features model global and long-range dependencies, offering context and overall structural information. The SSIM module dynamically fuses the outputs from phase, amplitude, and frequency interactions to produce enriched spectral-spatial tokens. LGM is used to extract rich



(a)

(b)

(c)

(d)

Fig. 11. Comparison of different training samples. (a) Xuzhou dataset. (b) ZYHHK dataset. (c) GF-5 dataset. (d) IP dataset.

semantic feature and modulate the local global information. The experiments result shows that the proposed model obtained satisfactory classification results in frequency domain.

Data availability

Salinas, Indian Pines Dataset available on https://www.ehu.eus/ccwintco/index.php? title=Hyperspectral_Remote_Sensing_Scenes. Xuzhou dataset available on https://ieee-dataport.org/documents/xuzhou-hyspex-dataset. The other datasets are available from the corresponding author on reasonable request.

Received: 30 May 2025; Accepted: 17 July 2025
Published online: 26 July 2025

References

1. Camps-Valls, G., Tuia, D., Bruzzone, L. & Benediktsson, J. A. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE. Signal. Process. Mag.* **31** (1), 45–54. https://doi.org/10.1109/MSP.2013.2279179 (2014).
2. Murphy, R., Whelan, B., Chlingaryan, A. & Sukkarieh, S. Quantifying leaf-scale variations in water absorption in lettuce from hyperspectral imagery: a laboratory study with implications for measuring leaf water content in the context of precision agriculture, *Precis. Agric.* **20**. https://doi.org/10.1007/s11119-018-9610-5 (2019).
3. Murphy, R., Chlingaryan, A. & Melkumyan, A. Predicting wavelength position of the ferric iron absorption at 900 nm from hyperspectral data (1000–2500 nm) (2013).
4. Haut, J. M. & Paoletti, M. E. Cloud implementation of multinomial logistic regression for UAV hyperspectral images. *IEEE J. Miniaturization Air Space Syst.* **1** (3), 163–171. https://doi.org/10.1109/JMASS.2020.3019669 (2020).
5. SahIn, Y. E., Arisoy, S. & Kayabol, K. Anomaly detection with Bayesian Gauss Background Model in hyperspectral images, in *26th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU.2018.8404293 (2018).
6. Chen, Y. N., Thaipisutikul, T., Han, C. C., Liu, T. J. & Fan, K. C. Feature line embedding based on support vector machine for hyperspectral image classification. *Remote Sens.* **13** (1). https://doi.org/10.3390/rs13010130 (2021).
7. Licciardi, G., Marpu, P. R., Chanussot, J. & Benediktsson, J. A. Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geosci. Remote Sens. Lett.* **9** (3), 447–451. https://doi.org/10.1109/LGRS.2011.2172185 (2012).

8. Villa, A., Benediktsson, J. A., Chanussot, J. & Jutten, C. Hyperspectral image classification with independent component discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **49** (12), 4865–4876. https://doi.org/10.1109/TGRS.2011.2153861 (2011).
9. Fauvel, M., Chanussot, J., Benediktsson, J. A. & Sveinsson, J. R. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles, in *IEEE International Geoscience and Remote Sensing Symposium*, 4834–4837. https://doi.org/10.1109/IGARSS.2007.4423943 (2007).
10. Dalla Mura, M., Villa, A., Benediktsson, J. A., Chanussot, J. & Bruzzone, L. Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. *IEEE Geosci. Remote Sens. Lett.* **8** (3), 542–546. https://doi.org/10.1109/LGRS.2010.2091253 (2011).
11. Zhao, W. & Du, S. Spectral–Spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote Sens.* **54** (8), 4544–4554. https://doi.org/10.1109/TGRS.2016.2543748 (2016).
12. Li, S. et al. Deep Learning for Hyperspectral Image Classification: An Overview, in *IEEE Transactions on Geoscience and Remote Sensing*, 1–20. https://doi.org/10.1109/TGRS.2019.2907932 (2019).
13. Chen, Y., Lin, Z., Zhao, X., Wang, G. & Gu, Y. Deep Learning-Based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **7** (6), 2094–2107. https://doi.org/10.1109/JSTARS.2014.2329330 (2014).
14. Chen, Y., Zhao, X. & Jia, X. Spectral–Spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **8** (6), 2381–2392. https://doi.org/10.1109/JSTARS.2015.2388577 (2015).
15. Hu, W., Huang, Y., Wei, L., Zhang, F. & Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.*, vol. p. 258619, Jul. 2015, (2015). https://doi.org/10.1155/2015/258619
16. Pan, B. et al. DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification. *IEEE Geosci. Remote Sens. Lett.* **17** (11), 1968–1972. https://doi.org/10.1109/LGRS.2019.2960528 (2020).
17. Song, W., Li, S., Fang, L. & Lu, T. Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote Sens.* **56** (6), 3173–3184. https://doi.org/10.1109/TGRS.2018.2794326 (2018).
18. Roy, S. K., Krishna, G., Dubey, S. R. & Chaudhuri, B. B. Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **17** (2), 277–281. https://doi.org/10.1109/LGRS.2019.2918719 (2020).
19. Tahir Arshad, J. Z. & Ullah, I. A hybrid Convolution transformer for hyperspectral image classification. *Eur. J. Remote Sens.* **2024**, 2330979. https://doi.org/10.1080/22797254.2024.2330979 (2024).
20. Paoletti, M. E. et al. Capsule networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **57** (4), 2145–2160. https://doi.org/10.1109/TGRS.2018.2871782 (2019).
21. Li, W., Wu, G., Zhang, F. & Du, Q. Hyperspectral image classification using deep Pixel-Pair features. *IEEE Trans. Geosci. Remote Sens.* **55** (2), 844–853. https://doi.org/10.1109/TGRS.2016.2616355 (2017).
22. Chen, Y. et al. Hyperspectral images classification with Gabor filtering and convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **14** (12), 2355–2359. https://doi.org/10.1109/LGRS.2017.2764915 (2017).
23. Aptoula, E., Ozdemir, M. C. & Yanikoglu, B. Deep learning with attribute profiles for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **13** (12), 1970–1974. https://doi.org/10.1109/LGRS.2016.2619354 (2016).
24. Mei, S., Ji, J., Hou, J., Li, X. & Du, Q. Learning Sensor-Specific Spatial-Spectral features of hyperspectral images via convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **55** (8), 4520–4533. https://doi.org/10.1109/TGRS.2017.2693346 (2017).
25. Li, J. et al. Classification of hyperspectral imagery using a new fully convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **15** (2), 292–296. https://doi.org/10.1109/LGRS.2017.2786272 (2018).
26. Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A. & Li, J. Visual Attention-Driven hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **57** (10), 8065–8080. https://doi.org/10.1109/TGRS.2019.2918080 (2019).
27. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *CoRR*, abs/2010.11929. https://arxiv.org/abs/2010.11929 (2020).
28. Chowdhary, K. R. Natural Language processing, in Fundamentals of Artificial Intelligence, 603–649. (Springer, New Delhi, 2020). https://doi.org/10.1007/978-81-322-3972-7_19.
29. Hong, D. et al. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers, *CoRR*, abs/2107.02988, 2021. https://arxiv.org/abs/2107.02988
30. He, X., Chen, Y. & Lin, Z. Spatial-Spectral transformer for hyperspectral image classification. *Remote Sens.* **13** (3). https://doi.org/10.3390/rs13030498 (2021).
31. Roy, S. K., Manna, S., Song, T. & Bruzzone, L. Attention-Based adaptive Spectral–Spatial kernel ResNet for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **59** (9), 7831–7843. https://doi.org/10.1109/TGRS.2020.3043267 (2021).
32. Roy, S. K. et al. Spectral–Spatial morphological attention transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15. https://doi.org/10.1109/TGRS.2023.3242346 (2023).
33. Zhao, Z., Xu, X., Li, S. & Plaza, A. Hyperspectral image classification using groupwise separable convolutional vision transformer network. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–17. https://doi.org/10.1109/TGRS.2024.3377610 (2024).
34. Meng, Z. et al. Global–Local multigranularity transformer for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **18**, 112–131. https://doi.org/10.1109/JSTARS.2024.3491294 (2025).
35. Meng, Z., Zhao, F. & Liang, M. A novel Spectral-Spatial MLP architecture for hyperspectral image classification. *Remote Sens.* **13** (20). https://doi.org/10.3390/rs13204060 (2021).
36. Fan, S., Liu, Q., Li, W. & Bai, H. A frequency and topology interaction network for hyperspectral image classification. *Eng. Appl. Artif. Intell.* **133**, 108234. https://doi.org/10.1016/j.engappai.2024.108234 (2024).
37. Sun, L., Zhao, G., Zheng, Y. & Wu, Z. Spectral–Spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14. https://doi.org/10.1109/TGRS.2022.3144158 (2022).
38. Ma, C. et al. Light Self-Gaussian-Attention vision transformer for hyperspectral image classification. *IEEE Trans. Instrum. Meas.* **72**, 1–12. https://doi.org/10.1109/TIM.2023.3279922 (2023).
39. Yang, X., Cao, W., Lu, Y. & Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15. https://doi.org/10.1109/TGRS.2022.3171551 (2022).
40. Bin Li Er, L. Z., Hu, O. W., Zhang, G. & Wu, J. Multi-granularity vision transformer via semantic token for hyperspectral image classification. *Int. J. Remote Sens.* **43** (17), 6538–6560. https://doi.org/10.1080/01431161.2022.2142078 (2022).
41. Ouyang, E. et al. When multigranularity Meets Spatial–Spectral attention: A hybrid transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–18. https://doi.org/10.1109/TGRS.2023.3242978 (2023).
42. Tang, X. et al. Hyperspectral image classification based on 3-D octave Convolution with Spatial–Spectral attention network. *IEEE Trans. Geosci. Remote Sens.* **59** (3), 2430–2447. https://doi.org/10.1109/TGRS.2020.3005431 (2021).
43. Qiao, X. & Huang, W. A dual frequency transformer network for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **16**, 10344–10358. https://doi.org/10.1109/JSTARS.2023.3328115 (2023).
44. Shi, H., Zhang, Y., Cao, G. & Yang, D. MHCFormer: multiscale hierarchical Conv-Aided fourierformer for hyperspectral image classification. *IEEE Trans. Instrum. Meas.* **73**, 1–15. https://doi.org/10.1109/TIM.2023.3344142 (2024).
45. Jun Yue, S. M., Zhao, W. & Liu, H. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* **6** (6), 468–477. https://doi.org/10.1080/2150704X.2015.1047045 (2015).
46. Xu, Y., Du, B. & Zhang, L. Beyond the patchwise classification: Spectral-Spatial fully convolutional networks for hyperspectral image classification. *IEEE Trans. Big Data.* **6** (3), 492–506. https://doi.org/10.1109/TBDATA.2019.2923243 (2020).

47. Chen, Y., Jiang, H., Li, C., Jia, X. & Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **54** (10), 6232–6251. https://doi.org/10.1109/TGRS.2016.2584107 (2016).
48. Rao, M., Tang, P. & Zhang, Z. A developed Siamese CNN with 3D adaptive Spatial-Spectral pyramid pooling for hyperspectral image classification. *Remote Sens.* **12** (12). https://doi.org/10.3390/rs12121964 (2020).
49. Wang, D., Du, B. & Zhang, L. Fully contextual network for hyperspectral scene parsing. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16. https://doi.org/10.1109/TGRS.2021.3050491 (2022).
50. Ghasrodashti, E. K., Adibi, P., Karshenas, H., Kashani, H. B. & Chanussot, J. Multimodal image classification based on convolutional network and Attention-Based hidden Markov random field. *IEEE Trans. Geosci. Remote Sens.* **63**, 1–14. https://doi.org/10.1109/TGRS.2025.3560913 (2025).
51. Li, X. et al. Mar., Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds, *Int. J. Netw. Dyn. Intell.* **2** (1), 93–116. https://doi.org/10.53941/ijndi0201006 (2023).
52. Huang, W., Zhu, Y. & Huang, R. Low light image enhancement network with attention mechanism and retinex model. *IEEE Access.* **8**, 74306–74314. https://doi.org/10.1109/ACCESS.2020.2988767 (2020).
53. Galassi, A., Lippi, M. & Torroni, P. Attention in natural Language processing. *IEEE Trans. Neural Networks Learn. Syst.* **32**, 4291–4308 (2019).
54. Chi, L., Jiang, B. & Mu, Y. Fast Fourier Convolution, in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 4479–4488 (Curran Associates, 2020). [Online]. https://proceedings.neurips.cc/paper_files/paper/2020/file/2fd5d41ec6cfab47e32164d5624269b1-Paper.pdf (2020).
55. Pan, X., Zang, C., Lu, W., Jiang, G. & Sun, Q. FSFF-Net: A Frequency-Domain feature and Spatial-Domain feature fusion network for hyperspectral image classification. *Electronics* **14** (11). https://doi.org/10.3390/electronics14112234 (2025).
56. He, S., Tian, J., Hao, L., Zhang, S. & Tian, Q. Unleashing the full potential of hyperspectral imaging: decoupled image and frequency-domain spatial–spectral framework. *Expert Syst. Appl.* **243**, 122870. https://doi.org/10.1016/j.eswa.2023.122870 (2024).
57. Shi, H. et al. F3Net: fast fourier filter network for hyperspectral image classification. *IEEE Trans. Instrum. Meas.* **72**, 1–18. https://doi.org/10.1109/TIM.2023.3277100 (2023).
58. Ilya, T. et al. O., MLP-Mixer: An all-MLP Architecture for Vision, vol. 2105.01601, 24261–24272 (2021).
59. Hugo Touvron and Piotr Bojanowski and Mathilde Caron and Matthieu Cord and Alaaeldin El-Nouby and Edouard Grave and Gautier Izacard and Armand Joulin and Gabriel Synnaeve and Jakob Verbeek and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training, [Online]. (2021). Available: https://arxiv.org/abs/2105.03404
60. Liu, H., Dai, Z., So, D. R. & Le, Q. V. Pay Attention to MLPs, in *Neural Information Processing Systems* [Online]. (2021). Available: https://api.semanticscholar.org/CorpusID:234742218
61. Lee-Thorp, J., Ainslie, J., Eckstein, I. & Ontañón, S. FNet: Mixing Tokens with Fourier Transforms, *ArXiv*, vol. abs/2105.03824, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:234336004
62. Rao, Y., Zhao, W., Zhu, Z., Lu, J. & Zhou, J. Global Filter Networks for Image Classification, *ArXiv*, vol. abs/2107.00645, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235694359
63. Guibas, J. et al. Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers, *ArXiv*, vol. abs/2111.13587, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:244709538
64. Ahmad, M. et al. Artifacts of different dimension reduction methods on hybrid CNN feature hierarchy for hyperspectral image classification. *Optik* **246**, 167757. https://doi.org/10.1016/j.ijleo.2021.167757 (2021).

Acknowledgements

This work is supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia through the Researchers Supporting Project (PNURSP2025R333).

Author contributions

Each author in this article contributed their distinct expertise and responsibilities in a collaborative way T.A. and B.P. were primarily responsible for formulating the study's concepts and determining the research direction. S.K. provided methodological frameworks so that the study's approach was rigorous and coherent. T.A and A.R. assumed responsibility for the software implementation, which is critical for data analysis and interpretation. T.A., S.K, and S.A. collaborated to validate the findings, ensuring the strength and reliability of the conclusions reached. N.A. supervised formal analytic techniques, while R.K. and S.K. carried out essential studies for data collection and interpretation. B.P. Supervised the use of resources, while S.K. thoroughly examined the collected data. S.K. handled further modifications and editing after T.A. initially wrote the manuscript. B.P. used data visualization to improve the presentation of significant findings. B.P and R.K. Provided supervision throughout the study process, ensuring complete conformity to scholarly guidelines. N.A. and S.A. collaborated on project administration tasks, with N.A. leading the funding acquisition task. All authors provided input during the manuscript drafting stage.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.U.k.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.