

Article

Self-Supervised Hyperspectral Image Clustering via Spatial–Frequency Interaction and Amplitude–Phase Decoupling

Heng Yuan ¹, Nan Huang ², Qichao Liu ³, Pengfei Liu ^{1,*}, Kang Ni ¹ and Zhizhong Zheng ²

¹ School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; b23041026@njupt.edu.cn (H.Y.); tznikang@njupt.edu.cn (K.N.)

² School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; huangnan@njupt.edu.cn (N.H.); 20220160@njupt.edu.cn (Z.Z.)

³ School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; liuqichao@njupt.edu.cn

* Correspondence: liupengfei@njupt.edu.cn

Highlights

What are the main findings?

- We proposed a self-supervised Spatial–Frequency Interaction and Amplitude–Phase Decoupling (SFI-APD) framework for hyperspectral image clustering.
- FreqViT performs query–key cross-spectrum modulation without constructing a dense token affinity matrix, reducing token aggregation complexity from (N^2D) to ($ND \log N$). SFI-APD achieves OAs of 57.47%, 79.38%, 54.57%, and 64.11% on the Indian Pines, Salinas, Pavia University, and Yangzhou datasets.

What is the implication of the main finding?

- The proposed framework provides an efficient alternative to quadratic self-attention for large-scale hyperspectral images while retaining long-range contextual modeling capability.
- The experimental and ablation results demonstrate that amplitude recalibration and phase preservation mitigate illumination-induced spectral variations, improve spatial boundary continuity, and enhance clustering robustness across diverse hyperspectral scenes.

Abstract

Hyperspectral image (HSI) clustering assigns unlabeled pixels to land-cover groups by jointly exploiting spectral and spatial observations. Existing Vision Transformer-based deep clustering captures global dependencies through self-attention. However, the quadratic computational complexity of self-attention restricts practical applications in large HSI scenes. Furthermore, illumination variation and topographic shading shift spectral amplitude of co-class pixels toward divergent directions in feature space, enlarging intra-class distances and reducing inter-class separability in learned embeddings. To address the above limitations, we propose a self-supervised Spatial–Frequency Interaction and Amplitude–Phase Decoupling framework, termed SFI-APD, which integrates a High-Order Spatial–Frequency Interaction Module (HSFIM), a Frequency Feature Attention Block (FFAB), and a Frequency-Domain Vision Transformer (FreqViT) into a unified architecture. Specifically, HSFIM couples local convolutions with Fourier filtering to extract enriched spectral–spatial representations. FFAB then decouples amplitude and phase components to suppress brightness variations, yielding illumination-robust embeddings. Finally, FreqViT performs attention modulation across spectral channels, reducing token aggregation complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$. On the Indian Pines, Salinas, Pavia University, and Yangzhou datasets, SFI-APD achieves OAs of 57.47%, 79.38%, 54.57%, and 64.11%, respectively, outperforming state-of-the-art self-supervised methods for large HSIs.



Academic Editor: Salah Bourennane

Received: 18 June 2026

Revised: 27 July 2026

Accepted: 30 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: hyperspectral image clustering; multi-domain fusion; frequency-domain feature attention; amplitude–phase decoupling; Vision Transformer

1. Introduction

Hyperspectral imagery (HSI) acquires spectral signatures of surface objects over hundreds of contiguous and narrow spectral bands, providing highly discriminative information for identifying land-cover categories with fine-grained spectral differences [1,2]. With such rich spectral characterization, HSI has been extensively applied in precision agriculture, mineral exploration, environmental monitoring, and other remote sensing tasks [3–5]. As a fundamental task in hyperspectral image interpretation, clustering aims to assign pixels into semantically coherent groups according to intrinsic spectral–spatial similarities without requiring manual annotations. However, the inherent high-dimensional redundancy of HSI and spectral distortions caused by illumination and topographic factors jointly limit the feature discriminability and spatial consistency of HSI clustering.

In recent years, CNN-based deep clustering has been developed for HSI clustering because convolutional encoders can extract local spectral–spatial continuity from neighborhood patches. DS³C-Net designs multi-scale autoencoders with pixel blocks of different spatial sizes and collaborative self-expressive layers to learn spatial–spectral subspace structures for HSI clustering [6]. Hu et al. map augmented HSI pairs into CNN feature space and optimize contrastive constraints before spectral clustering to improve separability among spatial–spectral representations [7]. DCAES²C-Net uses a convolutional autoencoder and an asymmetric autoencoder to extract spatial–spectral features and suppress shading- and texture-related non-material information before over-clustering-based k -means [8]. 3D-ACAE introduces a 3-D attention convolutional autoencoder and a clustering layer to produce compact spectral–spatial embeddings for HSI clustering [9]. CNN-based clustering methods preserve neighborhood texture continuity and local spectral consistency, but convolutional aggregation is still limited by kernel sizes and patch ranges. As a result, non-local relations among spatially separated regions with similar spectra remain difficult to model.

Transformer-based clustering models have recently been introduced into HSI clustering to compensate for limited receptive fields in CNN-based encoders. SDST constructs a self-supervised double-structure Transformer by combining an autoencoder-based autoformer branch with a Siamese dual-former graph branch, enabling multilevel feature fusion for global HSI representation learning [10]. STSDC couples a multiscale CNN block with a Semantic Transformer; semantic contiguous attention selects tokens from intra- and adjacent superpixels, and semantic global attention aggregates long-range superpixel representations to associate spatially separated land-cover regions [11]. Transformer-based HSI clustering studies indicate self-attention can supplement local convolutional encoders with non-local relation modeling. Standard self-attention, however, requires an $N \times N$ affinity matrix for N tokens, leading to $\mathcal{O}(N^2D)$ computational and memory costs. Efficient ViT designs therefore restrict attention to local windows or replace dot-product attention with kernelized approximations. Swin Transformer computes self-attention in shifted non-overlapping local windows to reduce image-scale attention cost [12]. CSWin Transformer computes self-attention in horizontal and vertical cross-shaped windows to enlarge token interaction ranges under reduced cost [13]. Window operations reduce computation, but direct interactions are removed for tokens located in different windows; homogeneous land-cover regions crossing window borders must rely on shifted or stacked layers for information exchange. Performer approximates softmax attention with positive

orthogonal random features to reduce attention computation to linear complexity [14]. Linear Transformer expresses self-attention as a linear dot product of kernel feature maps and uses matrix-product associativity to compute attention in linear form [15]. Kernelized formulations improve scalability, but approximation errors and feature-map compression may weaken fine boundary cues and long-range relations required by HSI clustering.

Local-window attention and linear attention reduce computational complexity by restricting token interactions or approximating dense affinity matrices. Window-based attention computes direct query–key relations inside local windows, so tokens located in different windows depend on shifted partitions or stacked layers for information exchange. Linear attention maps query and key features into kernel spaces, reducing the cost of explicit token-pair affinity computation. However, restricted window partitions and compressed query–key similarities may weaken long-range relations among spatially separated HSI regions with similar spectra, motivating alternative efficient token aggregation mechanisms beyond local partitioning and kernelized approximation.

Frequency-domain and wavelet-based ViT designs provide another route for efficient token aggregation [16–18]. Fourier-domain token mixing reduces dependence on dense token-pair affinity matrices, and wavelet-based attention preserves high-frequency details during key–value downsampling. However, existing frequency-domain and wavelet-based ViT designs do not explicitly separate illumination-sensitive amplitude responses from structure-related phase information. In HSIs, terrain-induced illumination differences and cast-shadow effects change spectral brightness for identical land-cover materials, increasing intra-class spectral variance. Inspired by low-light image enhancement [19], amplitude components are regarded as energy-related terms sensitive to illumination intensity, and phase components preserve edge locations and spatial geometry [20]. Coupled amplitude–phase processing may introduce illumination-biased energy responses into clustering embeddings and reduce boundary contrast between adjacent land-cover regions with similar spectra.

To address illumination-induced spectral variability, insufficient boundary continuity, and high attention cost in HSI clustering, we propose a self-supervised spatial–frequency interaction and amplitude–phase decoupling framework, termed SFI-APD. As shown in Figure 1, SFI-APD contains three components: a High-Order Spatial–Frequency Interaction Module (HSFIM), a Frequency Feature Attention Block (FFAB), and a Frequency-Domain Vision Transformer (FreqViT). HSFIM combines 3×3 convolution with learnable Fourier filtering. Local convolution extracts neighborhood texture cues, and Fourier filtering provides full-image spatial aggregation through element-wise spectral modulation. FFAB decomposes complex frequency features into amplitude and phase spectra. Amplitude recalibration adjusts illumination-sensitive energy responses, and phase preservation retains boundary-related geometric cues during reconstruction. FreqViT maps query, key, and value tokens into the frequency domain and uses query–key cross-spectrum modulation to aggregate long-range token relations. Frequency-domain token mixing reduces attention complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$ and provides efficient full-image contextual aggregation for HSI clustering.

The main contributions are summarized as follows:

- A self-supervised HSI clustering framework, termed SFI-APD, is proposed to learn clustering-oriented representations through spatial–frequency interaction and amplitude–phase decoupling. The High-Order Spatial–Frequency Interaction Module (HSFIM) combines local convolutional texture extraction with Fourier-domain global aggregation, improving boundary continuity and long-range consistency in clustering maps.

- FFAB is designed to decompose frequency features into amplitude and phase spectra. Amplitude recalibration adjusts illumination-sensitive energy responses caused by terrain-induced brightness variation, and phase preservation retains boundary-related spatial geometry during feature reconstruction.
- FreqViT is introduced for efficient long-range token aggregation in HSI clustering. Query–key correlation is computed through cross-spectrum modulation in the frequency domain, reducing standard attention complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$.

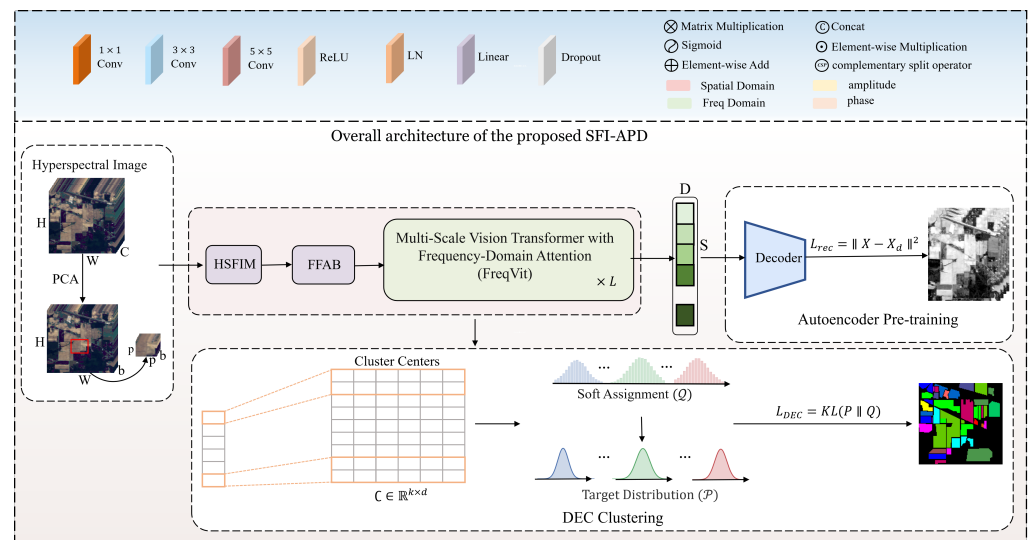


Figure 1. Overall architecture of the SFI-APD framework. Firstly, the High-order space–frequency interactive module (HSFIM) is used to fuse the space–frequency features of the hyperspectral image after PCA dimension reduction. Then Frequency Feature Attention Block (FFAB) decouples the features fused with spatial–frequency information from amplitude and phase to suppress the spectral distortion caused by illumination and retain structural information. Finally, the Frequency Domain Vision Transformer (FreqViT) captures long-range dependencies under a self-supervised optimization strategy for feature representation learning with clustering optimization.

2. Related Work

2.1. Deep Clustering

HSI clustering assigns unlabeled pixels to land-cover groups according to spectral–spatial similarity. HSI data often contain redundant spectral bands, mixed pixels, and similar spectra from different land-cover classes, making conventional clustering sensitive to noisy bands and local variations. Recent deep clustering studies can be grouped into autoencoder-based deep clustering, deep subspace clustering, graph-based deep clustering, and contrastive learning-based deep clustering.

Autoencoder-based deep clustering: Autoencoder-based deep clustering methods learn low-dimensional embeddings and cluster assignments through joint representation learning. DEC optimizes a KL-divergence clustering loss in latent space by iteratively refining soft assignments [21]. IDEC adds reconstruction loss to DEC, preserving local structure and clustering compactness during optimization [22]. DCN incorporates a k -means objective into representation learning to obtain clustering-oriented embeddings [23]. VaDE combines a variational autoencoder with a Gaussian mixture prior and estimates cluster assignments from latent posterior probabilities [24]. Autoencoder-based methods reduce spectral redundancy, but pixel-wise reconstruction alone provides limited spatial consistency for HSI clustering.

Deep subspace clustering: Deep subspace clustering methods combine deep embeddings with self-expressive constraints for affinity learning. DSC-Net inserts a self-expressive layer between encoder and decoder, learning a coefficient matrix for spectral clustering [25]. A full sample–sample coefficient matrix brings high storage and computation cost for large HSI scenes. EDESC replaces dense sample-wise self-expression with learnable subspace bases, reducing complexity from quadratic to linear with respect to sample number [26]. DSASC employs ViT features and separates structure/content representations for attention-aware subspace clustering [27]. Deep subspace clustering can model low-dimensional structures, but large-scale HSI clustering remains costly when affinity learning depends on many pixel pairs.

Graph-based deep clustering: Graph-based deep clustering methods construct pixel or superpixel graphs and learn embeddings through graph neural networks. DGAE builds spatial and spectral similarity graphs from superpixels and extracts graph embeddings through dual graph autoencoders [28]. IcicleGCN combines instance-level contrast, cluster-level learning, and multiscale neighborhoods in a GCN-based clustering model [29]. MLGSC constructs spectral and texture graph views and applies multilevel contrastive losses to learn local–global graph embeddings [30]. Graph-based methods exploit spatial adjacency, but clustering quality depends on graph construction and graph computation becomes expensive for large HSI scenes.

Contrastive learning-based deep clustering: Contrastive learning-based deep clustering methods learn embeddings by increasing agreement between positive pairs and reducing similarity between negative pairs. SSCC uses a symmetric twin network and a projection head for spectral–spatial augmentations, improving intra-cluster compactness and inter-cluster separation [31]. SDCGSC couples superpixel segmentation with dual-neighborhood graph autoencoders to preserve local spectral information and broader spatial context [32]. CMSCGC constructs textural and spectral–spatial graph convolution views and uses contrastive learning plus attention fusion to build an affinity matrix for spectral clustering [33]. Contrastive methods improve embedding separability, but positive/negative pair construction and graph-view design affect clustering stability across different HSI scenes.

2.2. Vision Transformer

Transformers use multi-head self-attention to project tokens into query, key, and value spaces, compute scaled query–key similarities, and aggregate value tokens according to attention weights [34]. ViT extends Transformers to image recognition by splitting images into patches, mapping patches into token embeddings, and processing token sequences with Transformer encoders [35]. Global self-attention builds an $N \times N$ affinity matrix for N tokens, leading to $\mathcal{O}(N^2D)$ computation and memory costs. Large HSI scenes contain many spatial pixels and spectral bands, so dense token interaction increases memory consumption and training time for HSI clustering.

Transformer-based HSI clustering methods have been developed to model non-local spectral–spatial relations. SDST combines an autoformer branch with a Siamese dual-former graph branch, fusing multilevel features for self-supervised HSI clustering [10]. STSDC integrates a multiscale CNN block with a Semantic Transformer, using semantic contiguous attention and semantic global attention for large HSI clustering [11]. Transformer-based clustering methods improve non-local relation modeling, but standard self-attention still brings quadratic complexity.

Efficient ViT designs reduce attention cost through window-based attention or linearized attention. Swin Transformer computes attention inside shifted local windows, reducing computation and enabling cross-window information exchange through shifted

partitions [12]. CSWin Transformer performs attention along horizontal and vertical cross-shaped windows, enlarging token interaction range with controlled cost [13]. Window-based attention reduces memory consumption, but tokens in different windows cannot interact directly within a single attention layer. Linear attention methods approximate dense attention through kernel feature maps or random features, reducing complexity to near-linear form [14,15]. Other efficient attention and token-mixing variants further replace explicit dense attention with cosine reweighting, attention-free updates, Fourier token mixing, global filtering, DCT decorrelation, or frequency-domain self-attention [36–41]. Approximation and window partitioning can weaken long-range relations among separated HSI regions with similar spectra.

Frequency-domain and wavelet-based ViT architectures present an alternative paradigm for efficient token mixing. AFNO employs adaptive Fourier-domain token mixing for high-resolution Vision Transformers via block-diagonal channel mixing, adaptive weight sharing, and sparse frequency modes [16]. Wave-ViT integrates wavelet transforms into Vision Transformer attention, leveraging invertible key–value downsampling to retain high-frequency texture details [17]. SpectFormer combines multi-head attention with spectral layers that transform image tokens into the Fourier domain and recover spatial tokens through inverse Fourier transforms [18]. Earlier Fourier token mixers, such as FNet, substitute self-attention with parameter-free Fourier transforms, whereas GFNet formulates global frequency filters to capture long-range spatial dependencies [38,39]. DCT-based attention leverages frequency decorrelation and high-frequency component truncation to mitigate computational overhead, whereas FFTformer-FSAS approximates scaled dot-product attention via element-wise multiplication in the frequency domain [40,41].

Rather than relying on fixed spectral mixing or rigid frequency filtering, FreqViT computes input-adaptive query–key correlations via normalized cross-spectrum modulation. This avoids constructing a dense $N \times N$ affinity matrix, successfully preserving token dependencies while reducing aggregation complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$. Furthermore, while conventional frequency-domain models target generic vision tasks, FreqViT explicitly integrates amplitude–phase decoupling. The design allows the model to capture long-range spectral–spatial dependencies and robustly handle illumination variations in self-supervised HSI clustering.

3. Method

3.1. Overall Architecture

As illustrated in Figure 1, SFI-APD is designed for self-supervised HSI clustering by combining spatial–frequency interaction, amplitude–phase decoupling, and frequency-domain token aggregation. Given an input HSI feature cube $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote spatial height, spatial width, and channel number, respectively, SFI-APD contains three main modules: High-Order Spatial–Frequency Interaction Module (HSFIM), Frequency Feature Attention Block (FFAB), and Frequency-Domain Vision Transformer (FreqViT).

HSFIM extracts local texture responses through convolution and models global spatial aggregation through learnable Fourier filtering. Spatial and frequency responses are fused by multiplicative interaction and residual addition, producing the spatial–frequency feature $\mathbf{F}_{sf}$. FFAB decomposes $\mathbf{F}_{sf}$ into amplitude and phase components in the Fourier domain. In Fourier analysis, amplitude and magnitude both refer to the complex modulus; magnitude is used in equations for notational clarity. The amplitude branch recalibrates channel-wise energy responses related to illumination variation, and the phase branch keeps boundary-related angular information during complex-spectrum reconstruction. FreqViT then converts the FFAB output into token sequences and performs query–key

cross-spectrum modulation in the frequency domain. Finally, a self-supervised clustering objective jointly optimizes reconstruction loss and KL-divergence clustering loss to obtain pixel-wise cluster assignments.

3.2. High-Order Spatial–Frequency Interaction Module

HSI clustering requires both neighborhood texture cues and long-range spatial relations. HSFIM uses parallel spatial and frequency branches to extract complementary responses from $\mathbf{X}$. The spatial branch applies a 3×3 convolution to capture local texture and edge cues. The frequency branch maps $\mathbf{X}$ into the Fourier domain, modulates Fourier coefficients by a learnable complex filter, and transforms the filtered coefficients back to the spatial domain. The two branches are formulated as

$$\begin{aligned} \mathbf{F}_{\text{spa}} &= \phi_{3 \times 3}(\mathbf{X}), \\ \mathbf{F}_{\text{freq}} &= \text{Re} \left[\mathcal{F}_{2\text{D}}^{-1}(\mathbf{W}_c \odot \mathcal{F}_{2\text{D}}(\mathbf{X})) \right], \end{aligned} \quad (1)$$

where $\phi_{3 \times 3}(\cdot)$ denotes a 3×3 convolution, $\mathcal{F}_{2\text{D}}(\cdot)$ and $\mathcal{F}_{2\text{D}}^{-1}(\cdot)$ denote 2-D FFT and 2-D IFFT, $\mathbf{W}_c \in \mathbb{C}^{H \times W \times C}$ denotes a learnable complex filter, $\odot$ denotes the Hadamard product, and $\text{Re}[\cdot]$ extracts the real part. According to the convolution theorem, element-wise multiplication in the Fourier domain corresponds to circular filtering in the spatial domain. Therefore, $\mathbf{F}_{\text{freq}}$ provides full-image aggregation without constructing dense pixel-wise affinity matrices. The detailed architecture of HSFIM is shown in Figure 2.

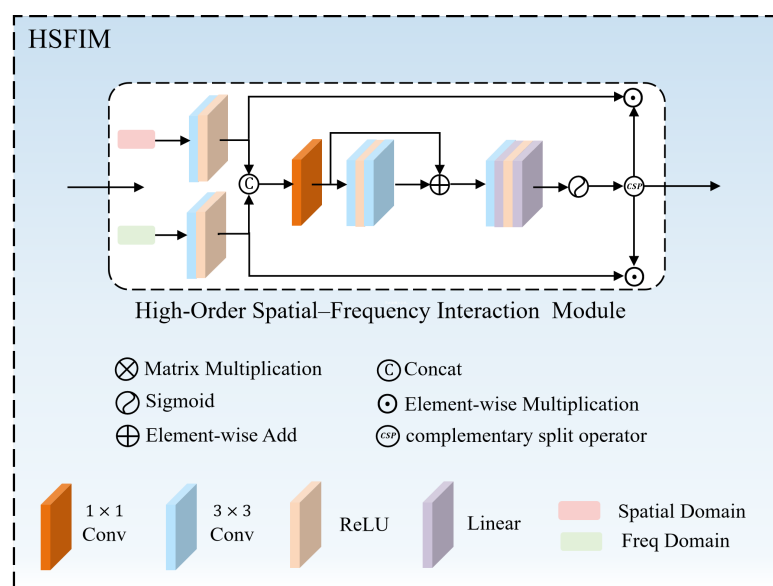


Figure 2. Detailed architecture of HSFIM.

The spatial and frequency responses are fused as

$$\mathbf{F}_{\text{sf}} = \phi_{1 \times 1}(\mathbf{F}_{\text{spa}} \odot \mathbf{F}_{\text{freq}}) + \mathbf{F}_{\text{spa}} + \mathbf{F}_{\text{freq}}, \quad (2)$$

where $\phi_{1 \times 1}(\cdot)$ denotes a 1×1 convolution. In Equation (2), $\mathbf{F}_{\text{freq}}$ reweights local responses from $\mathbf{F}_{\text{spa}}$, and residual addition keeps branch-specific information for subsequent amplitude–phase processing.

3.3. Frequency Feature Attention Block

The HSFIM output $\mathbf{F}_{\text{sf}}$ contains convolutional texture responses and Fourier-filtered global responses. Illumination variation mainly changes Fourier amplitude, whereas phase angles are closely related to spatial layout and edge positions. FFAB therefore processes

amplitude and phase components separately. The block contains two steps: amplitude–phase decomposition and amplitude recalibration with phase preservation. The detailed architecture of FFAB is shown in Figure 3.

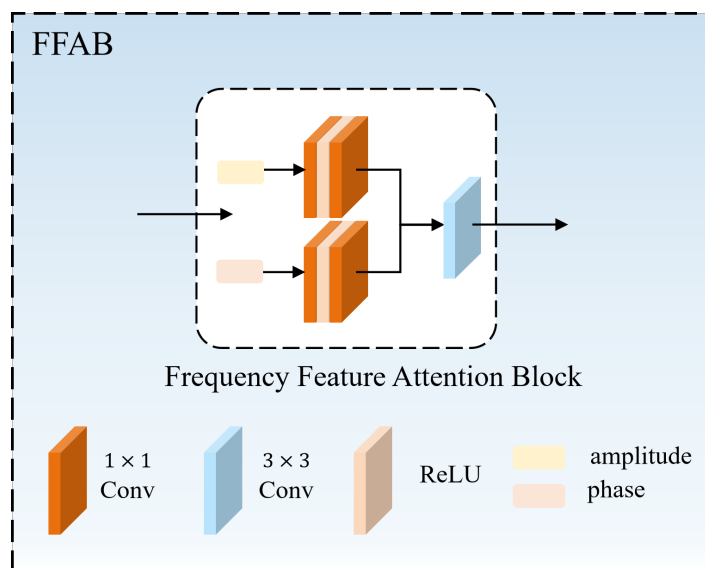


Figure 3. Detailed architecture of FFAB.

3.3.1. Amplitude–Phase Decomposition

FFAB first applies 2-D FFT to $\mathbf{F}_{sf}$ and obtains a complex spectrum $\mathbf{S}_{sf}$. The amplitude tensor $\mathbf{M}_{sf}$ and phase tensor Φ_{sf} are computed by complex modulus and phase angle:

$$\begin{aligned}\mathbf{S}_{sf} &= \mathcal{F}_{2D}(\mathbf{F}_{sf}), \\ \mathbf{M}_{sf} &= |\mathbf{S}_{sf}|, \\ \Phi_{sf} &= \angle(\mathbf{S}_{sf}),\end{aligned}\quad (3)$$

where $|\cdot|$ denotes the complex modulus, and $\angle(\cdot)$ denotes the phase angle. The amplitude tensor $\mathbf{M}_{sf}$ records coefficient energy at different spatial frequencies and channels. Low-frequency amplitude components are associated with brightness and illumination intensity. The phase tensor Φ_{sf} records angular information of Fourier coefficients, which is related to spatial layout and boundary positions.

3.3.2. Amplitude Recalibration and Phase Preservation

Amplitude recalibration uses channel-wise statistics from $\mathbf{M}_{sf}$. Global average pooling summarizes amplitude responses along spatial–frequency coordinates:

$$\mathbf{z}_m(c) = \frac{1}{HW} \sum_{u=1}^H \sum_{v=1}^W \mathbf{M}_{sf}(u, v, c), \quad c = 1, \dots, C. \quad (4)$$

The vector $\mathbf{z}_m \in \mathbb{R}^C$ is passed into an MLP and a sigmoid function to generate a channel-wise amplitude gate:

$$\mathbf{g}_m = \text{Reshape}[\sigma(\text{MLP}(\mathbf{z}_m))], \quad \mathbf{g}_m \in [0, 1]^{1 \times 1 \times C}. \quad (5)$$

The amplitude tensor is recalibrated by residual scaling, and the phase tensor is retained:

$$\begin{aligned}\hat{\mathbf{M}}_{sf} &= \mathbf{M}_{sf} \odot (\mathbf{1} + \mathbf{g}_m), \\ \hat{\Phi}_{sf} &= \Phi_{sf}.\end{aligned}\quad (6)$$

Equation (6) adjusts channel-wise amplitude responses and keeps phase angles for boundary-related spatial geometry.

The updated amplitude and phase are combined into a reconstructed complex spectrum:

$$\hat{\mathbf{S}}_{\text{sf}} = \hat{\mathbf{M}}_{\text{sf}} \odot \exp(j\hat{\Phi}_{\text{sf}}), \quad (7)$$

where j denotes the imaginary unit. The reconstructed spectrum is mapped back to the spatial domain. A 1×1 projection mixes channel responses after inverse Fourier transform, and residual fusion retains spatial–frequency information from HSFIM:

$$\mathbf{F}_{\text{ap}} = \phi_{1 \times 1} \left(\text{Re} \left[\mathcal{F}_{2D}^{-1} \left(\hat{\mathbf{S}}_{\text{sf}} \right) \right] \right) + \mathbf{F}_{\text{sf}}. \quad (8)$$

3.4. Frequency-Domain Vision Transformer

FFAB produces $\mathbf{F}_{\text{ap}}$, which contains recalibrated amplitude responses and phase-preserved spatial geometry. FreqViT further models long-range token relations through frequency-domain cross-spectrum modulation. Standard self-attention computes $\mathbf{QK}^T$ and requires $\mathcal{O}(N^2D)$ operations for N tokens and embedding dimension D . FreqViT replaces dense token-pair affinity with element-wise multiplication of Fourier-transformed queries and keys. The detailed architecture of FreqViT is shown in Figure 4.

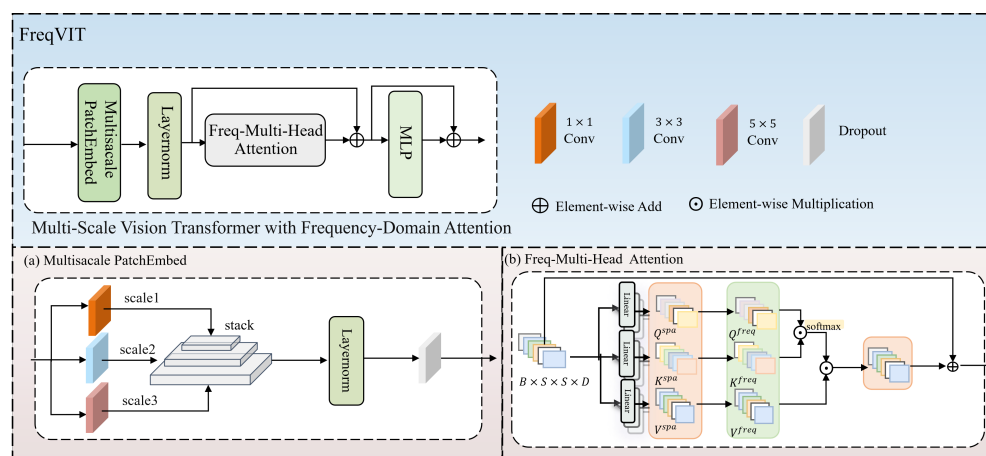


Figure 4. Detailed architecture of FreqViT.

3.4.1. Multi-Scale Patch Embedding

The feature map $\mathbf{F}_{\text{ap}}$ is converted into tokens through multi-scale patch embedding. Parallel overlapping convolution branches with kernel sizes $\{P, P/2, P/4\}$ extract local regions at different spatial scales:

$$\mathbf{T} = \text{Linear}(\text{Concat}[\psi_P(\mathbf{F}_{\text{ap}}), \psi_{P/2}(\mathbf{F}_{\text{ap}}), \psi_{P/4}(\mathbf{F}_{\text{ap}})]), \quad (9)$$

where $\psi_P(\cdot)$, $\psi_{P/2}(\cdot)$, and $\psi_{P/4}(\cdot)$ denote overlapping convolutional embeddings with different kernel sizes. After flattening, $\mathbf{T} \in \mathbb{R}^{N \times D}$ denotes the token sequence. Learnable position encoding $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times D}$ is added to retain spatial order:

$$\mathbf{T}_e = \mathbf{T} + \mathbf{E}_{\text{pos}}. \quad (10)$$

3.4.2. Frequency-Domain Self-Attention

The embedded token sequence is projected into query, key, and value tensors:

$$\begin{aligned}\mathbf{Q} &= \mathbf{T}_e \mathbf{W}_Q, \\ \mathbf{K} &= \mathbf{T}_e \mathbf{W}_K, \\ \mathbf{V} &= \mathbf{T}_e \mathbf{W}_V,\end{aligned}\quad (11)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices. A 1-D RFFT along the token dimension maps $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into the frequency domain:

$$\begin{aligned}\mathbf{Q}_f &= \mathcal{F}_{1D}^r(\mathbf{Q}), \\ \mathbf{K}_f &= \mathcal{F}_{1D}^r(\mathbf{K}), \\ \mathbf{V}_f &= \mathcal{F}_{1D}^r(\mathbf{V}),\end{aligned}\quad (12)$$

where $\mathcal{F}_{1D}^r(\cdot)$ denotes 1-D RFFT. According to the cross-correlation theorem, sequence-domain correlation can be represented by conjugate multiplication in the frequency domain. FreqViT computes the query-key cross-spectrum and generates a real-valued frequency attention map:

$$\mathbf{A}_f = \text{Softmax}_{N_f} \left(\frac{|\mathbf{Q}_f \odot \text{conj}(\mathbf{K}_f)|}{\sqrt{d_h}} \right). \quad (13)$$

Specifically, for each attention head, the frequency-domain tensors $\mathbf{Q}_f$, $\mathbf{K}_f$, and $\mathbf{V}_f$ are of shape $B \times h \times N_f \times d_h$, where B denotes the batch size, h is the number of attention heads, d_h represents the channel dimension per head, and $N_f = \lfloor N/2 \rfloor + 1$ corresponds to the frequency dimension derived from the token sequence of length N . The Softmax_f operation is applied independently along the frequency axis for each batch element, attention head, and feature channel. Subsequently, the normalized query-key cross-spectrum is utilized to modulate the corresponding frequency components of the value tensor $\mathbf{V}_f$.

The frequency attention map modulates the value spectrum, and 1-D IRFFT restores modulated tokens to the sequence domain:

$$\mathbf{Z} = \text{Linear} \left((\mathcal{F}_{1D}^r)^{-1} (\mathbf{A}_f \odot \mathbf{V}_f) \right). \quad (14)$$

FreqViT reduces token aggregation complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$ and does not store an $N \times N$ affinity matrix.

3.5. Self-Supervised Learning Module

The self-supervised learning module jointly optimizes reconstruction and clustering objectives. The encoder maps input patches into latent embeddings $\{\mathbf{z}_i\}_{i=1}^n$, and K cluster centers are denoted as $\{\boldsymbol{\mu}_j\}_{j=1}^K$. Following DEC [21], Student's t -distribution computes the soft assignment probability between $\mathbf{z}_i$ and $\boldsymbol{\mu}_j$:

$$q_{ij} = \frac{\left(1 + \|\mathbf{z}_i - \boldsymbol{\mu}_j\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'=1}^K \left(1 + \|\mathbf{z}_i - \boldsymbol{\mu}_{j'}\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}. \quad (15)$$

Parameter α controls the degrees of freedom. The soft cluster frequency f_j and target distribution p_{ij} are defined as

$$\begin{aligned}f_j &= \sum_{i=1}^n q_{ij}, \\ p_{ij} &= \frac{q_{ij}^2 / f_j}{\sum_{j'=1}^K q_{ij'}^2 / f_{j'}}.\end{aligned}\quad (16)$$

The reconstruction loss constrains the decoded output $\hat{\mathbf{X}}$ to preserve information from the input $\mathbf{X}$:

$$\mathcal{L}_{\text{rec}} = \frac{1}{n} \sum_{i=1}^n \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2. \quad (17)$$

The clustering loss is defined by KL divergence between the target distribution $\mathbf{P} = [p_{ij}]$ and the soft assignment distribution $\mathbf{Q} = [q_{ij}]$:

$$\mathcal{L}_{\text{clu}} = \sum_{i=1}^n \sum_{j=1}^K p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (18)$$

The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{clu}}, \quad (19)$$

where λ controls the contribution of the clustering loss. The reconstruction term preserves spectral-spatial information in latent embeddings, and the KL term pulls high-confidence samples toward corresponding cluster centers.

4. Experiments

Experiments use four hyperspectral image datasets: Indian Pines (IP), Salinas (SA), Pavia University (PU), and Yangzhou (YZ). SFI-APD is compared with representative clustering methods from autoencoder-based clustering, graph clustering, deep subspace clustering, and contrastive clustering. Quantitative metrics, class-wise producer's accuracy (PA), clustering maps, ablation studies, and attention-mechanism comparisons are reported.

4.1. Experimental Setup

4.1.1. Datasets

Indian Pines (IP). IP was acquired by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over an agricultural area in northwestern Indiana, USA. The image contains 145×145 pixels with a spatial resolution of 20 m. After removal of water-absorption and noisy bands, 200 spectral bands are retained. The reference map contains 16 land-cover classes [42].

Salinas (SA). SA was acquired by AVIRIS over Salinas Valley, CA, USA. The image contains 512×217 pixels with a spatial resolution of 3.7 m. After removal of water-absorption bands, 204 spectral bands are used. The reference map contains 16 land-cover classes [43].

Pavia University (PU). PU was acquired by Reflective Optics System Imaging Spectrometer (ROSIS) over Pavia, Italy. The image contains 610×340 pixels with a spatial resolution of 1.3 m. After removal of noisy bands, 103 spectral bands are retained from 115 original bands covering 0.43–0.86 μm . The reference map contains 9 land-cover classes [44].

Yangzhou (YZ). YZ was acquired over Yangzhou, China, by Multi-Parameter Airborne Remote Sensing System (MARSS). Data collection was performed at a flight altitude of 600 m, and spatial resolution is 1.5 m. The image contains 956×365 pixels, 249 continuous spectral bands, and 8 land-cover classes [45].

4.1.2. Evaluation Metrics

Quantitative evaluation uses overall accuracy (OA), adjusted Rand index (ARI), adjusted mutual information (AMI), normalized mutual information (NMI), Homogeneity, Completeness, V-measure, and Fowlkes–Mallows index (FMI). Class-level evaluation uses PA. OA is computed after optimal one-to-one matching between predicted clusters and reference classes.

Compared methods: Nine representative methods are selected for comparison. DCN [23], IDEC [22], and DEKM [46] are autoencoder-based clustering methods. CCGC [47] is a graph clustering method. DSASC [27] and EDESC [26] are deep subspace clustering methods. SLCGC [48] and SeCu [49] are contrastive clustering methods. In addition, SDST [10], a self-supervised double-structure Transformer proposed for HSI clustering in 2024, is included as a recent Transformer-based state-of-the-art baseline. In particular, SDST combines an autoformer branch with a Siamese dual-former graph branch to perform multilevel spectral–spatial feature fusion. Together with SLCGC, which was published in 2025, the comparison includes recent task-aligned methods from the past two years. The selected baselines cover reconstruction-driven representation learning, graph-structured relation modeling, subspace affinity learning, contrastive representation learning, and Transformer-based global dependency modeling.

4.1.3. Implementation Details

All datasets are first reduced by principal component analysis (PCA) before network training. Candidate PCA dimensions are selected from $\{10, 20, 30, 40, 50\}$, and candidate patch sizes are selected from $\{3, 5, 7, 9, 11\}$. Based on the parameter selection results, the PCA dimension is set to 30 and the patch size is set to 5 for all datasets. The cluster number is set to reference-class number in each dataset.

The training is performed in two consecutive stages. In Stage 1, the autoencoder is pretrained to minimize the reconstruction loss using the Adam optimizer with a learning rate of 1×10^{-3} . The number of pretraining rounds for the IP, SA, PU, and YZ datasets is set to 200, 300, 600, and 800, respectively. Stage 1 initializes the encoder and decoder weights and enables the network to learn preliminary feature representations with spectral–spatial structure. In Stage 2, the encoder and decoder weights pretrained in Stage 1 are used to initialize the clustering network. The encoder extracts latent features from the input data, and k-means is applied to initialize the cluster centers. The DEC-based clustering algorithm then iteratively updates the latent feature representation and cluster centers by minimizing the KL divergence between the target distribution P and the soft assignment distribution Q , with the degree-of-freedom parameter in Student's t -distribution fixed at $\alpha = 1$ for all datasets. The clustering loss weight is set to 1.0, and the learning rate is set to 1×10^{-4} . All experiments were implemented using PyTorch version 2.1.0+cu121 on an NVIDIA GeForce RTX 4090 GPU.

4.2. Comparison with State-of-the-Art Methods

Tables 1–8 report quantitative comparisons on four datasets. Figures 5–8 present clustering maps. Bold values denote highest metric values.

4.2.1. Results on Indian Pines

Tables 1 and 2 list class-wise PA and global metrics on IP. Figure 5 presents clustering maps. IP contains many crop categories with close spectral responses, mixed field boundaries, and several rare classes.

Table 1 presents the class-wise results, reflecting high Producer's Accuracy (PA) and balanced User's Accuracy (UA) for SFI-APD. Hay-windrowed reaches 100.00% PA, Grass-trees reaches 88.89% PA, Wheat reaches 86.79% PA, and Woods reaches 79.60% PA. In terms of UA, SFI-APD achieves 90.78% for Grass-pasture, 89.25% for Grass-trees, and 87.81% for Woods, indicating reliable predictions for the major vegetation categories. Rare samples in Alfalfa and Oats restrict the cluster-center estimation. Complementing the category-specific observations, the global quantitative evaluations further validate the overall structural consistency. Table 2 shows 57.47% OA, 47.06% AA, 52.01% Kappa, 39.71% ARI, 57.68% AMI, 57.88% NMI, and 46.83% FMI for SFI-APD. Among autoencoder-based baselines, DEKM

achieves the highest OA with 47.47%. Among graph and subspace baselines, DSASC gives the highest OA with 47.51%. Among contrastive baselines, SLCGC gives the highest OA with 45.27%. SFI-APD improves OA by 10.00% over DEKM, 12.69% over CCGC, 9.96% over DSASC, and 12.20% over SLCGC. SFI-APD also improves NMI by 7.08% over DEKM and 8.80% over SLCGC. The superior OA, Kappa, and NMI on the Indian Pines dataset validate the effectiveness of HSFIM and FFAB. HSFIM connects separated crop parcels with close spectra through spatial–frequency interaction. FFAB reduces illumination-related amplitude variation and keeps phase-related boundary geometry.

Table 1. Per-class producer’s accuracy (PA) and user’s accuracy (UA) comparison on the Indian Pines dataset (%).

Metric	Class	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
PA	Alfalfa	0.00	0.00	0.00	0.00	18.52	1.85	0.00	0.00	0.00	0.00
	Corn-notill	27.82	36.40	32.85	28.45	43.03	34.03	32.22	24.55	27.20	44.56
	Corn-mintill	47.00	35.37	32.37	26.74	24.94	42.69	41.49	5.76	62.11	64.39
	Corn	0.00	3.42	20.09	42.74	1.71	15.81	13.25	0.43	20.09	34.62
	Grass-pasture	47.69	64.99	60.36	64.14	30.38	63.18	65.39	0.00	61.57	65.39
	Grass-trees	70.82	84.74	85.81	92.64	86.61	70.28	85.14	38.82	49.80	88.89
	Grass-pasture-mowed	88.46	57.69	0.00	38.46	0.00	23.08	84.62	0.00	96.15	0.00
	Hay-windrowed	80.78	76.07	99.80	100.00	31.08	99.59	94.07	79.96	95.50	100.00
	Oats	100.00	0.00	0.00	90.00	0.00	0.00	0.00	0.00	0.00	5.00
	Soybean-notill	42.87	27.48	44.63	42.46	24.17	32.64	42.15	16.22	0.00	31.30
	Soybean-mintill	44.21	42.06	39.30	31.52	43.31	39.26	39.59	72.69	79.78	59.72
	Soybean-clean	12.54	19.22	18.08	19.22	45.60	29.48	14.66	21.34	11.73	11.89
	Wheat	39.15	84.91	94.34	86.32	62.74	74.06	98.11	70.75	96.23	86.79
	Woods	36.86	53.63	56.34	60.17	82.27	57.96	60.66	70.56	52.70	79.60
UA	Buildings-Grass-Trees-Drivers	28.42	19.47	12.63	60.69	69.21	5.53	31.84	23.95	42.89	28.16
	Stone-Steel-Towers	68.42	18.95	4.21	100.00	90.53	89.47	52.63	0.00	100.00	52.63
	Alfalfa	0.00	0.00	0.00	0.00	5.92	0.06	0.00	0.00	0.00	0.00
	Corn-notill	67.17	49.34	47.05	44.84	65.29	56.22	58.11	45.60	84.97	43.83
	Corn-mintill	26.70	13.79	18.28	20.86	13.76	46.97	16.18	10.91	31.07	39.43
	Corn	0.00	7.14	22.38	16.18	1.80	12.13	8.27	0.88	20.26	14.11
	Grass-pasture	31.77	56.08	42.43	96.31	86.78	40.36	47.72	0.00	64.97	90.78
	Grass-trees	68.08	97.38	94.26	91.29	78.33	81.90	93.12	47.93	100.00	89.25
	Grass-pasture-mowed	9.75	6.25	0.00	6.49	0.00	2.75	10.84	0.00	7.69	0.00
	Hay-windrowed	92.72	92.08	81.47	86.55	40.64	87.91	90.55	71.87	90.68	74.09
	Oats	5.32	0.00	0.00	3.86	0.00	0.00	0.00	0.00	0.00	0.55
	Soybean-notill	31.18	29.13	33.51	40.61	61.90	51.72	33.58	19.41	0.00	50.75
	Soybean-mintill	57.39	49.31	50.76	62.59	53.69	52.01	50.49	52.36	49.24	67.46
	Soybean-clean	30.80	22.74	19.30	23.79	21.88	27.30	28.48	21.83	40.22	40.78
	Wheat	41.09	67.67	60.79	76.57	55.88	42.90	88.51	17.44	59.82	66.19
	Woods	96.95	92.53	87.41	90.39	84.13	90.14	89.51	73.45	99.71	87.81
	Buildings-Grass-Trees-Drivers	21.30	66.67	17.58	17.13	51.87	14.69	50.42	13.94	29.80	45.73
	Stone-Steel-Towers	95.59	27.27	10.26	59.37	49.71	95.51	94.34	0.00	92.23	90.91

Note: Bold values indicate the highest value for each class and metric.

Table 2. Overall clustering performance on the Indian Pines dataset (%).

Metric	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
OA	41.59	43.95	45.45	44.78	47.51	45.27	47.47	41.66	51.23	57.47
AA	45.94	39.03	37.55	55.22	40.88	42.43	47.24	26.56	49.73	47.06
Kappa	35.02	36.85	38.80	39.46	41.19	39.43	40.93	33.56	43.88	52.01
ARI	24.03	26.42	26.69	24.01	30.90	28.11	28.32	27.98	31.77	39.71
AMI	46.53	46.77	47.31	47.08	47.46	48.85	50.56	35.52	55.90	57.68
NMI	46.78	47.02	47.56	47.34	47.71	49.08	50.80	35.78	56.12	57.88
Homogeneity	48.05	47.40	48.76	50.27	48.57	50.91	51.14	34.54	53.80	58.73
Completeness	45.57	46.65	46.42	44.73	46.87	47.39	50.45	37.12	58.65	57.05
V-measure	46.78	47.02	47.56	47.34	47.71	49.08	50.80	35.78	56.12	57.88
FMI	32.62	35.36	35.03	32.10	38.96	36.14	36.96	38.23	43.11	46.83

Note: Bold values indicate the highest value for each metric.

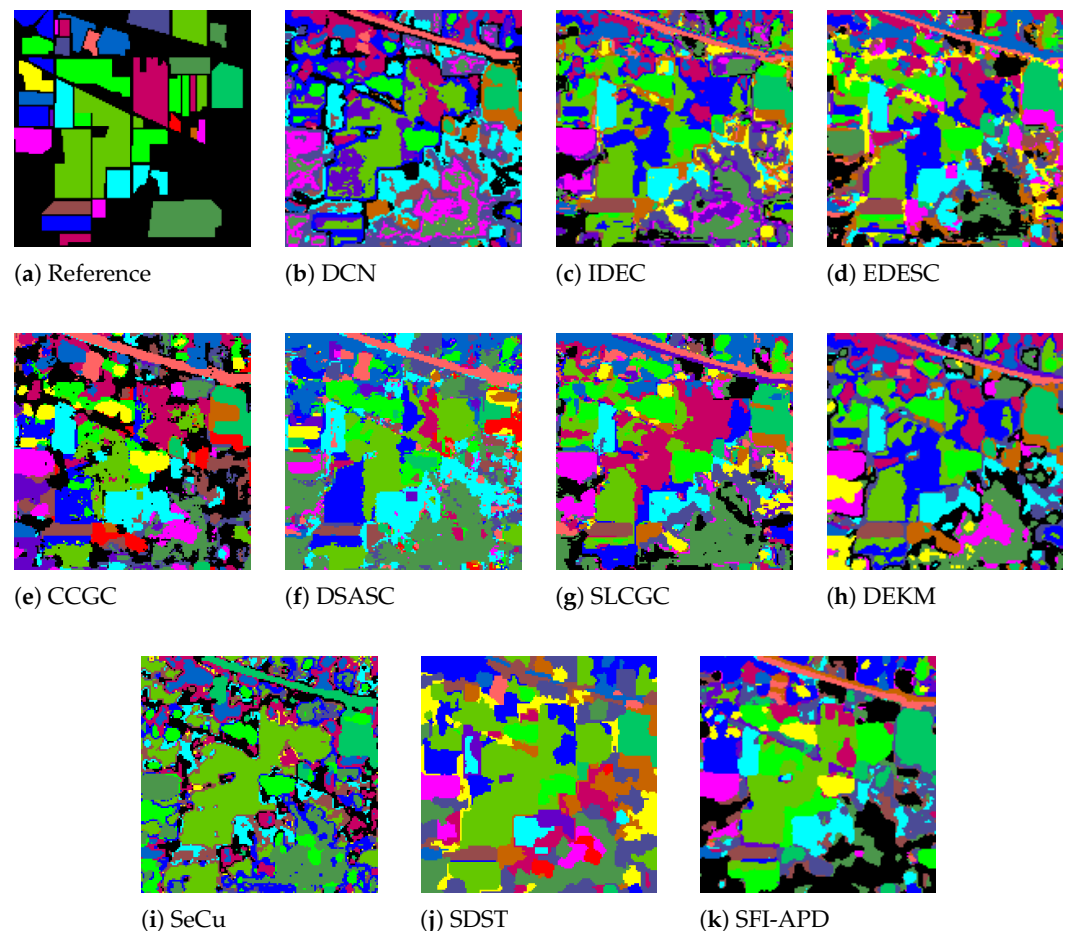


Figure 5. Clustering maps of the Indian Pines dataset.

4.2.2. Results on Salinas

Tables 3 and 4 report class-wise and global results on SA. Figure 6 shows clustering maps. SA contains regular agricultural parcels, and several crop categories or lettuce growth stages have close spectral signatures.

Table 3 presents the class-wise results, showing high completeness and reliable User's Accuracy (UA) for several large crop fields. SFI-APD achieves 100.00% PA for Broccoli green weeds 2 and Stubble, 99.60% for Fallow, 99.40% for Grapes untrained, and 99.79% for Soil vineyard develop. Regarding UA, SFI-APD reaches 100.00% for Broccoli green weeds 1 and at least 98.00% for Broccoli green weeds 2, Stubble, Celery, Soil vineyard develop, and Vineyard vertical trellis, demonstrating reliable predictions for several major crop categories. Class confusion occasionally occurs among Fallow rough plow, Lettuce romaine 4 wk, Lettuce romaine 6 wk, and Vineyard untrained. Lettuce growth stages share close reflectance and often appear in adjacent areas, leading to merged unsupervised cluster centers. Despite localized confusion among spectrally similar classes, the global evaluation metrics in Table 4 demonstrate the robust clustering capability of the proposed method. Table 4 shows 79.38% OA, highly competitive AA and Kappa coefficients, 72.99% ARI, 90.37% AMI, 90.37% NMI, and 78.33% FMI for SFI-APD. SLCGC gives the highest baseline OA with 73.82%. SFI-APD improves OA by 5.56%, NMI by 5.25%, and ARI by 4.51% over SLCGC. Among autoencoder-based baselines, IDEC reaches 68.89% OA, and DEKM reaches 82.58% NMI. Among subspace and graph baselines, EDESC gives 81.81% NMI, and CCGC gives 60.78% OA. Frequency-domain aggregation provides additional long-range grouping capabilities beyond conventional reconstruction, graph construction, and contrastive pairs.

Table 3. Per-class producer’s accuracy (PA) and user’s accuracy (UA) comparison on the Salinas dataset (%).

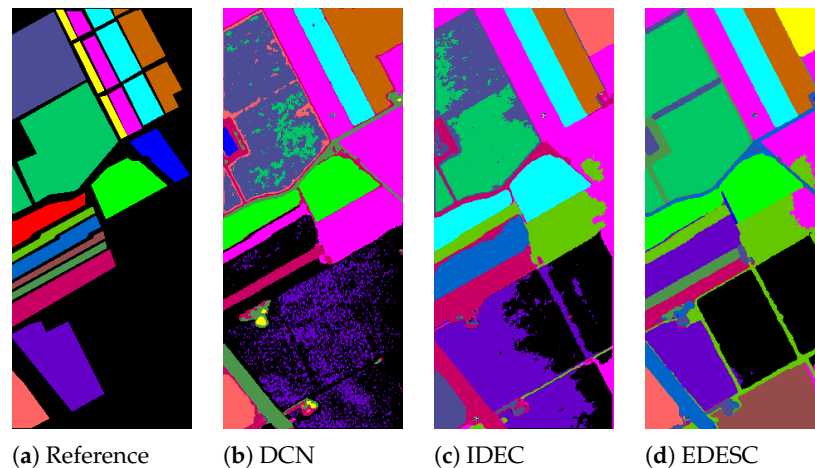
Metric	Class	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
PA	Broccoli green weeds 1	0.00	0.00	0.00	89.35	62.32	0.00	100.00	96.76	94.47	96.22
	Broccoli green weeds 2	99.97	0.00	100.00	96.22	53.27	99.49	99.97	84.43	100.00	100.00
	Fallow	0.05	0.00	0.00	28.14	99.95	99.70	0.00	45.95	99.60	99.60
	Fallow rough plow	0.00	0.00	0.00	0.00	92.47	0.14	0.00	0.00	98.78	0.00
	Fallow smooth	98.32	98.88	98.17	8.44	76.10	98.39	96.45	95.41	47.39	91.08
	Stubble	98.86	99.39	99.75	98.11	71.99	99.32	100.00	99.47	99.75	100.00
	Celery	99.78	99.72	100.00	96.48	100.00	99.64	100.00	99.66	99.69	99.64
	Grapes untrained	19.55	84.78	96.18	78.56	35.60	94.43	98.25	49.94	80.41	99.40
	Soil vineyard develop	69.95	99.61	99.19	59.68	99.00	99.15	100.00	98.00	99.98	99.79
	Corn senesced green weeds	34.35	97.38	62.90	86.77	69.77	93.56	40.00	82.52	89.44	92.53
	Lettuce romaine 4 wk	0.00	99.91	99.72	0.19	0.00	96.82	0.00	82.40	93.73	0.00
	Lettuce romaine 5 wk	8.98	76.49	0.00	3.74	73.48	35.91	0.05	26.98	0.00	85.00
	Lettuce romaine 6 wk	0.11	0.00	0.00	0.00	59.72	0.00	0.00	0.22	0.00	0.00
	Lettuce romaine 7 wk	0.00	0.00	23.64	0.00	0.00	96.92	0.00	0.00	91.78	98.50
UA	Vineyard untrained	89.96	78.03	2.88	59.48	51.22	0.00	0.00	62.64	0.78	0.00
	Vineyard vertical trellis	96.07	0.00	99.56	61.10	96.68	84.17	97.23	2.16	98.40	98.73
	Broccoli green weeds 1	0.00	0.00	0.00	52.44	56.86	0.00	100.00	51.40	100.00	100.00
	Broccoli green weeds 2	64.96	0.00	64.95	95.98	100.00	64.85	100.00	99.87	97.11	98.00
	Fallow	100.00	0.00	0.00	74.43	72.60	99.80	0.00	86.15	91.15	58.61
	Fallow rough plow	0.00	0.00	0.00	0.00	70.40	1.85	0.00	0.00	53.17	0.00
	Fallow smooth	36.93	47.50	47.87	6.57	63.59	64.25	36.61	56.82	95.99	90.97
	Stubble	100.00	40.69	99.50	44.50	100.00	99.19	98.29	89.16	99.80	99.27
	Celery	100.00	99.30	99.69	99.97	99.47	99.28	100.00	70.97	100.00	99.97
	Grapes untrained	91.30	85.76	60.98	75.78	75.46	60.37	60.52	72.72	55.57	60.82
	Soil vineyard develop	43.52	99.95	62.81	76.33	98.22	97.29	61.33	84.65	88.63	98.39
	Corn senesced green weeds	67.75	83.32	98.42	96.17	61.93	78.42	37.11	54.49	66.76	97.37
	Lettuce romaine 4 wk	0.00	66.27	67.15	2.17	0.00	71.46	0.00	52.10	34.16	0.00
	Lettuce romaine 5 wk	8.54	42.99	0.00	4.29	34.83	88.49	100.00	76.13	0.00	59.80
	Lettuce romaine 6 wk	100.00	0.00	0.00	0.00	40.85	0.00	0.00	0.41	0.00	0.00
	Lettuce romaine 7 wk	0.00	0.00	14.32	0.00	0.00	36.08	0.00	0.00	51.31	51.19
	Vineyard untrained	44.53	62.57	41.72	81.54	66.48	0.00	0.00	65.78	25.45	0.00
	Vineyard vertical trellis	58.51	0.00	100.00	29.46	99.49	99.67	100.00	1.67	96.89	99.72

Note: Bold values indicate the highest value for each class and metric.

Table 4. Clustering results of different methods on the Salinas dataset (%).

Metric	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
OA	55.35	68.89	67.00	60.78	64.37	73.82	66.88	67.37	73.47	79.38
AA	44.75	52.14	55.12	47.89	65.10	68.60	52.00	57.91	74.64	72.53
Kappa	51.80	64.36	60.56	55.98	60.95	70.39	63.28	63.67	70.34	76.78
ARI	50.22	64.29	62.94	50.71	51.99	68.48	62.87	53.90	65.73	72.99
AMI	74.05	80.03	81.79	70.83	74.54	85.10	82.75	71.97	84.77	90.37
NMI	74.08	80.04	81.81	70.86	74.56	85.12	82.58	71.99	84.78	90.37
Homogeneity	67.66	72.95	73.76	69.74	76.86	79.50	72.84	71.11	81.27	84.65
Completeness	81.84	88.67	91.83	72.02	72.39	91.59	95.34	72.89	88.61	96.93
V-measure	74.08	80.04	81.81	70.86	74.56	85.12	82.58	71.99	84.78	90.37
FMI	57.47	69.38	70.30	55.76	56.77	74.10	70.95	58.49	70.65	78.33

Note: Bold values indicate the highest value for each metric.

**Figure 6.** Cont.

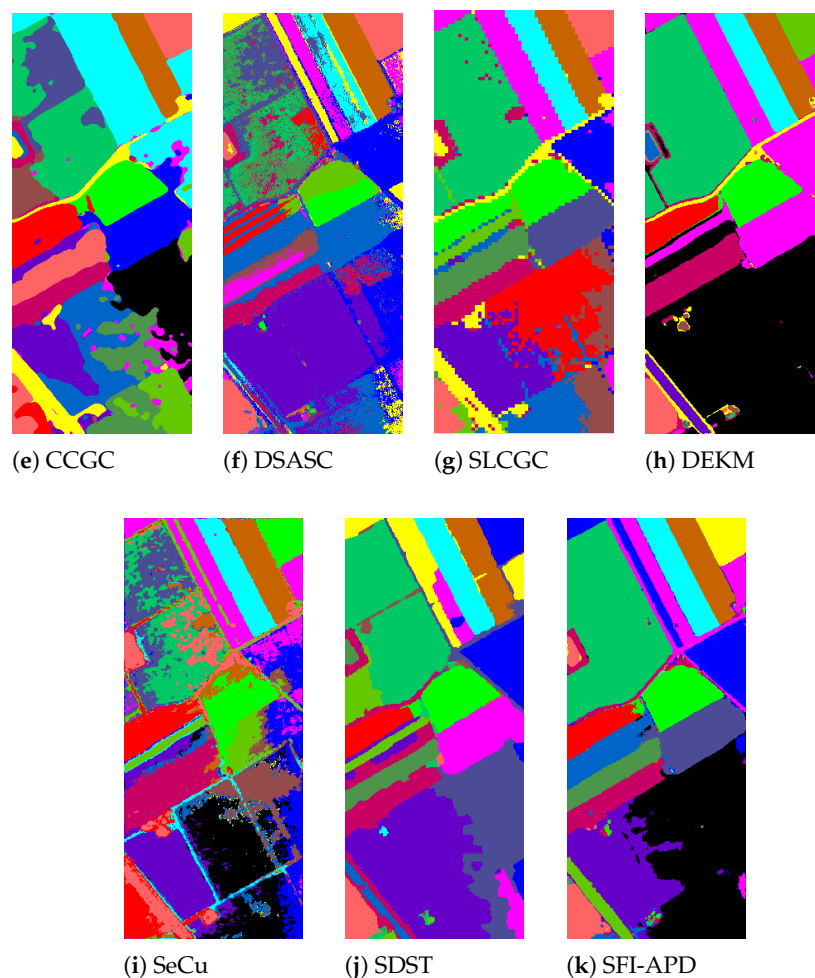


Figure 6. Clustering maps of the Salinas dataset.

4.2.3. Results on Pavia University

Tables 5 and 6 report results on PU. Figure 7 presents visual maps. PU is an urban scene with sharp boundaries, shadow regions, and land-cover materials with overlapping spectral signatures.

Table 5. Class-wise producer's accuracy (PA) and user's accuracy (UA) comparison on the Pavia University dataset (%).

Metric	Class	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
PA	Asphalt	96.21	82.19	57.08	52.68	82.90	72.33	35.94	43.95	57.99	55.96
	Meadows	56.06	42.63	36.82	58.81	25.22	38.65	5.34	50.96	43.53	53.06
	Gravel	40.50	81.99	87.04	0.14	4.00	0.00	82.90	79.23	68.70	0.43
	Trees	82.67	50.13	65.08	71.91	26.80	66.84	100.00	63.02	59.17	69.32
	Painted metal sheets	100.00	100.00	100.00	81.93	94.87	92.57	72.38	95.32	96.80	100.00
	Bare Soil	38.56	0.00	43.23	39.93	44.60	31.97	0.00	58.02	84.03	50.79
	Bitumen	0.00	0.00	0.00	17.82	0.00	0.15	74.96	0.30	0.00	0.00
	Self-Blocking Bricks	0.00	0.05	40.71	73.83	89.49	80.69	0.00	35.85	73.30	95.68
	Shadows	3.91	74.02	1.16	0.00	0.00	0.00	0.00	68.43	0.00	19.01
UA	Asphalt	42.06	37.41	58.49	64.69	68.73	65.02	92.74	69.15	58.67	58.39
	Meadows	86.97	76.34	72.90	84.11	91.14	79.81	41.79	97.52	94.95	88.45
	Gravel	50.87	36.06	33.17	0.19	1.83	0.00	50.30	32.49	55.21	0.13
	Trees	50.33	61.05	38.11	83.93	22.57	93.82	99.93	83.38	83.01	47.00
	Painted metal sheets	99.93	91.68	42.44	34.02	44.95	14.33	29.29	31.12	99.24	99.70
	Bare Soil	28.51	0.00	61.92	30.01	58.87	40.14	0.00	76.95	31.22	93.76
	Bitumen	0.00	0.00	0.00	12.38	0.00	0.14	43.66	0.08	0.00	0.00
	Self-Blocking Bricks	0.00	0.06	50.80	43.92	46.45	45.51	0.00	21.31	40.36	47.72
	Shadows	11.71	12.68	1.76	0.00	0.00	0.00	0.00	29.70	0.00	7.96

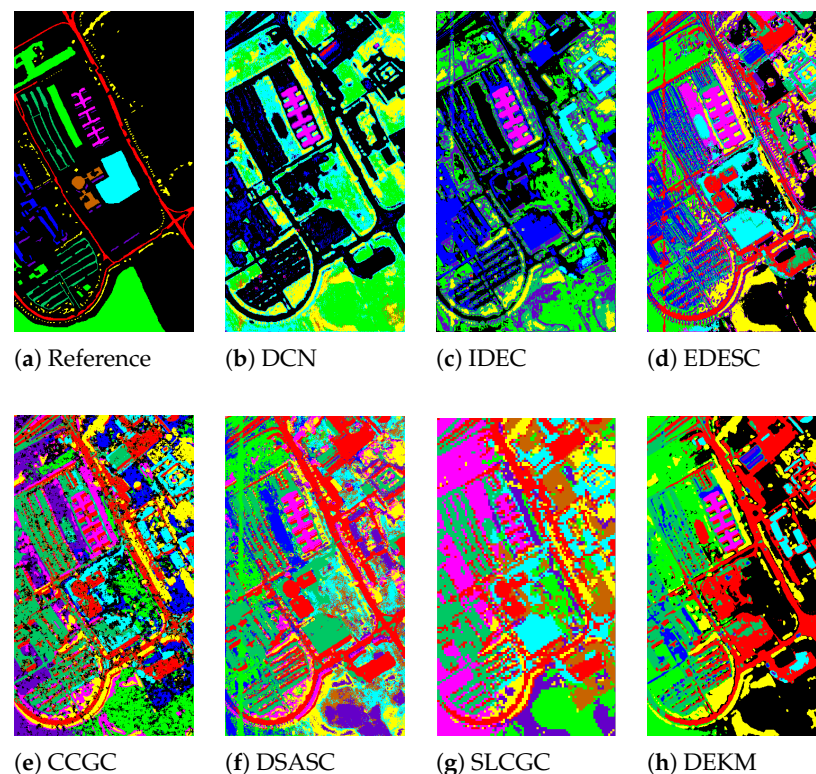
Note: Bold values indicate the highest value for each class and metric.

Table 6. Clustering results of different methods on the Pavia University dataset (%).

Metric	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
OA	55.03	43.73	45.59	46.13	41.89	46.47	52.70	51.86	54.81	54.57
AA	46.43	47.89	47.90	44.12	40.88	42.58	41.28	55.01	53.72	49.36
Kappa	41.08	32.15	39.90	55.98	30.68	36.49	39.00	42.02	45.54	44.31
ARI	28.96	18.27	25.11	27.91	25.06	31.56	32.23	32.16	37.82	39.40
AMI	46.79	40.34	46.94	45.80	50.79	51.39	52.02	45.79	55.95	53.26
NMI	46.81	40.36	46.96	45.83	50.81	51.41	52.04	45.81	55.96	53.28
Homogeneity	44.61	39.87	51.06	50.42	56.42	54.09	50.96	50.40	56.76	55.75
Completeness	49.25	40.87	43.47	42.00	46.21	49.89	53.17	41.99	55.19	51.02
V-measure	46.81	40.36	46.96	45.83	50.81	51.41	52.04	45.81	55.96	53.28
FMI	46.43	37.11	39.96	42.34	39.61	45.80	47.58	45.94	51.69	52.38

Note: Bold values indicate the highest value for each metric.

Table 5 presents the class-wise results, showing 100.00% PA for Painted metal sheets and 95.68% PA for Self-Blocking Bricks, while Bare Soil reaches 50.79% PA. For UA, SFI-APD achieves 99.70% for Painted metal sheets, 93.76% for Bare Soil, and 88.45% for Meadows, indicating high prediction precision for these urban categories. Multi-scale embedding describes repeated urban texture, and FFAB keeps boundary geometry during amplitude–phase reconstruction. Gravel, Bitumen, and Shadows show class confusion because road-like materials and shadowed surfaces have close spectral responses and local context. Beyond the analysis of specific urban materials, the global metrics provide a comprehensive evaluation of the overall clustering stability. Table 6 shows 54.57% OA, competitive AA and Kappa, 39.40% ARI, 53.26% AMI, 53.28% NMI, 53.28% V-measure, and 52.38% FMI for SFI-APD. DCN obtains the highest OA with 55.03% because matched clusters cover several dominant Asphalt and Meadows regions. SFI-APD obtains higher ARI and FMI than DCN by 10.44% and 5.95%, respectively, and maintains stable agreement in Kappa. Higher pairwise metrics indicate more consistent grouping for smaller urban materials and shadow-affected areas. SDST obtains the highest Homogeneity with 56.76%, and the corresponding 46.21% Completeness indicates the fragmentation of several reference classes. Overall, the balanced evaluation metrics confirm the capability of SFI-APD to maintain structural consistency in complex urban environments.

**Figure 7.** Cont.

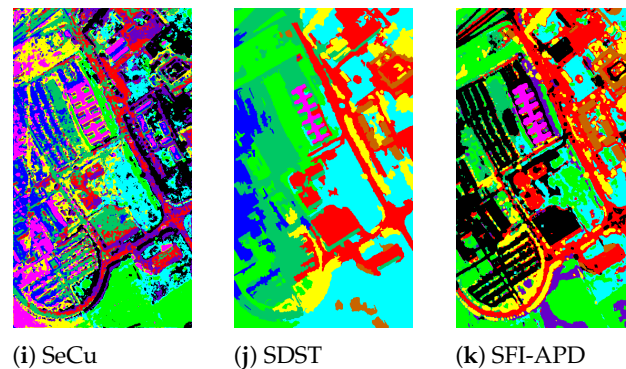


Figure 7. Clustering maps of the Pavia University dataset.

4.2.4. Results on Yangzhou

Tables 7 and 8 report class-wise and global results on YZ. Figure 8 presents clustering maps. YZ has larger spatial extent and more complex land-cover distribution than three public datasets.

Table 7. Class-wise producer’s accuracy (PA) and user’s accuracy (UA) comparison on the Yangzhou dataset (%).

Metric	Class	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST	SFI-APD
PA	luotu	1.04	30.89	60.23	51.17	61.64	56.84	2.58	45.78	58.84	72.33
	shuini	3.77	5.74	3.83	73.42	6.08	8.71	3.54	58.32	54.03	2.99
	shuizao	0.11	0.00	48.89	35.98	11.65	42.78	0.00	23.46	0.68	3.68
	jianzhu	62.75	98.22	70.34	5.13	59.99	49.40	99.70	10.49	29.31	0.00
	caodi	56.68	25.00	11.64	79.36	0.00	85.26	0.00	77.20	32.15	83.86
	tantu	83.11	56.25	32.75	87.43	28.84	4.11	64.30	97.48	0.00	0.00
	liqing	60.36	0.70	60.85	51.26	26.17	31.39	0.00	44.83	0.00	52.95
	zazao	0.00	0.00	0.00	34.47	10.00	68.84	0.00	48.63	0.28	90.33
UA	luotu	1.37	50.29	19.22	84.35	46.79	54.69	4.24	67.73	76.48	54.24
	shuini	1.38	6.71	0.42	34.71	0.08	2.97	1.46	42.72	75.04	4.28
	shuizao	0.03	0.00	2.36	2.93	1.41	5.47	0.00	6.99	0.01	1.03
	jianzhu	100.00	50.46	99.99	0.30	97.23	3.83	46.68	3.44	6.09	0.00
	caodi	47.90	0.12	6.48	96.96	0.00	98.72	0.00	69.20	1.04	98.92
	tantu	56.65	68.80	95.81	33.56	94.71	2.55	71.24	45.56	0.00	0.00
	liqing	62.18	1.18	66.49	92.59	72.67	91.41	0.00	87.80	0.00	81.91
	zazao	0.00	0.00	0.00	79.79	0.01	64.33	0.00	77.23	0.02	55.72

Note: Bold values indicate the highest value for each class and metric.

Table 8. Clustering results of different methods on the Yangzhou dataset (%).

Metric	DCN	IDEC	EDESC	CCGC	DSASC	SLCGC	DEKM	SeCu	SDST [†]	SFI-APD
OA	57.49	49.45	52.27	54.64	44.84	53.09	48.94	56.89	54.86	64.11
AA	33.48	27.10	36.07	52.28	25.55	43.42	21.27	50.77	30.31	38.27
Kappa	41.62	40.53	42.34	44.26	29.38	44.64	40.26	48.53	47.68	51.34
ARI	36.22	26.27	40.67	43.21	34.03	46.17	24.71	35.74	42.83	46.00
AMI	51.46	36.19	51.20	52.27	48.23	51.31	37.84	45.56	49.99	53.37
NMI	51.46	36.20	51.20	52.27	48.23	51.31	37.84	45.56	50.00	53.37
Homogeneity	50.02	33.26	57.79	59.92	59.66	57.08	31.83	50.63	59.09	55.46
Completeness	52.99	39.72	45.97	46.36	40.48	46.60	46.66	41.42	43.33	51.44
V-measure	51.46	36.20	51.20	52.27	48.23	51.31	37.84	45.56	50.00	53.37
FMI	53.31	48.05	52.65	54.86	46.97	57.29	49.89	48.56	54.40	57.69

Note: Bold values indicate the highest value for each metric.

To illustrate the performance on individual categories, Table 7 shows the quantitative clustering accuracy for the Yangzhou dataset. Table 7 reports high PA and balanced UA on luotu, caodi, and zazao. Decreased PA on shuini, shuizao, jianzhu, and tantu indicates spectral mixing between roads, buildings, tidal flats, and water-related surfaces. The complex spectral mixing explains the higher clustering difficulty of the Yangzhou dataset compared to agricultural scenes with regular parcels. Despite the inherent challenges of spectrally similar surfaces, the proposed method demonstrates strong global stability. From Table 8, SFI-APD achieves higher performances on the Yangzhou dataset compared with

the other representative baselines. Table 8 shows 64.11% OA, stable AA and Kappa, 53.37% AMI, 53.37% NMI, 53.37% V-measure, and 57.69% FMI for SFI-APD. Specifically, among autoencoder-based clustering methods, DCN gives the highest baseline OA with 57.49%. Compared with DCN, SeCu, and CCGC, SFI-APD incorporates advanced spatial–frequency aggregation to achieve better clustering performance. SFI-APD improves OA by 6.62%, 7.22%, and 9.47%, respectively. SFI-APD also improves NMI and FMI by 1.91% and 4.38% over DCN, confirming the capability to handle complex urban and water-related scenes.

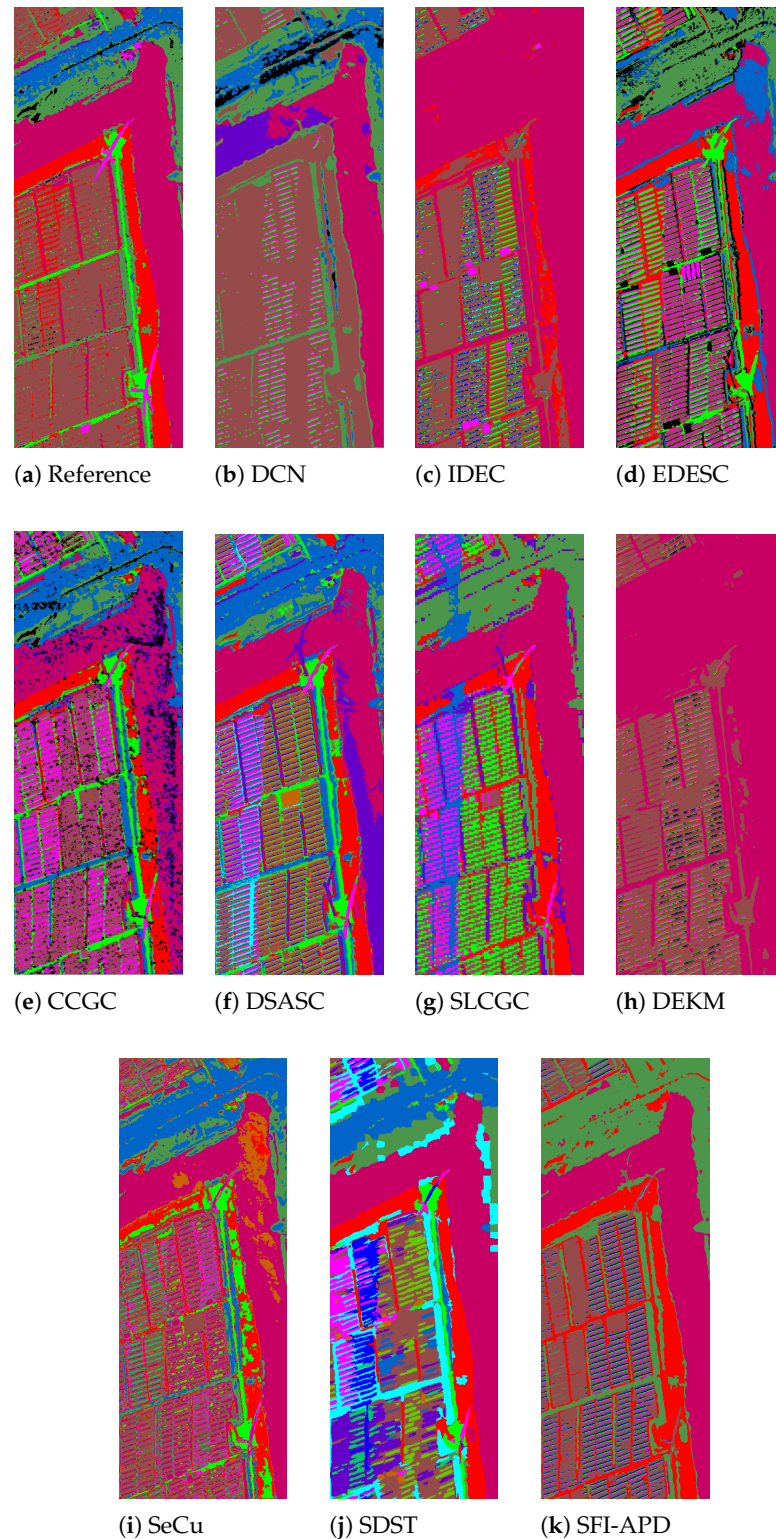


Figure 8. Clustering maps of the Yangzhou dataset.

4.3. Ablation Studies

Ablation studies evaluate contributions from HSFIM, FFAB, FreqViT, multi-scale convolutional embedding, and phase-related frequency attention variants. All ablation experiments use the same training strategy and data preprocessing as main comparison.

4.3.1. Effectiveness of Main Components

Table 9 reports module-level ablation results. SFI-APD-H removes FFAB. SFI-APD-A removes HSFIM. SFI-APD-V removes HSFIM and FFAB. SFI-APD-S replaces FreqViT with Spatial Attention. SFI-APD-M removes HSFIM, FFAB, and FreqViT. SFI-APD denotes a complete framework.

Table 9. Module-level ablation results of SFI-APD.

Method	HSFIM	FFAB	FreqViT	Spatial Attention	FLOPs (G)	Params (K)	IP OA (%)	SA OA (%)	PU OA (%)	YZ OA (%)
SFI-APD-H	✓	×	✓	×	70.647	191.451	51.24	76.03	51.83	57.56
SFI-APD-A	×	✓	✓	×	48.130	146.330	51.87	74.97	48.49	52.94
SFI-APD-V	×	×	✓	×	42.280	134.480	50.30	75.12	48.51	49.75
SFI-APD-S	✓	✓	×	✓	79.114	203.301	49.73	71.25	51.85	51.72
SFI-APD-M	×	×	×	×	44.897	134.480	51.23	77.18	47.92	51.85
SFI-APD	✓	✓	✓	×	76.497	203.301	57.47	79.38	54.57	64.11

Note: ✓ indicates that the corresponding module is used, whereas × indicates that it is not used. Bold values indicate the best results; lower values are better for FLOPs and Params, whereas higher values are better for OA.

Table 9 shows OA degradation after removing or replacing core modules. Removing FFAB reduces OA by 6.23%, 3.35%, 2.74%, and 6.55% on IP, SA, PU, and YZ, respectively. FFAB recalibrates frequency amplitude and preserves phase geometry, so removal increases illumination-related shifts and boundary confusion. Removing HSFIM reduces OA by 5.60%, 4.41%, 6.08%, and 11.17%. HSFIM introduces Fourier-domain global filtering into local convolutional responses, so removal weakens full-scene contextual aggregation.

The joint removal of HSFIM and FFAB reduces OA by 7.17%, 4.26%, 6.06%, and 14.36% on IP, SA, PU, and YZ, respectively. Replacing FreqViT with Spatial Attention reduces OA by 7.74%, 8.13%, 2.72%, and 12.39%. FreqViT computes query–key correlation through frequency-domain cross-spectrum modulation and avoids storage of dense token affinity. Component ablation therefore supports three module roles: HSFIM supplies local–global spatial representation, FFAB adjusts illumination-sensitive amplitude and retains phase geometry, and FreqViT performs efficient long-range token aggregation.

4.3.2. Effectiveness of Multi-Scale Convolutional Embedding

Table 10 compares single-scale 3×3 convolutional embedding with adopted multi-scale convolutional embedding.

Table 10. Ablation study on multi-scale convolutional embedding.

Method	IP OA (%)	SA OA (%)	PU OA (%)	YZ OA (%)
Single-scale 3×3 CNN	48.23	75.19	47.76	56.18
Multi-scale CNN	57.47	79.38	54.57	64.11

Note: Bold values indicate the highest OA on each dataset.

Table 10 indicates multi-scale embedding improves OA from 48.23% to 57.47% on IP, from 75.19% to 79.38% on SA, from 47.76% to 54.57% on PU, and from 56.18% to 64.11% on YZ. Corresponding gains are 9.24%, 4.19%, 6.81%, and 7.93%. Single-scale 3×3 convolution describes a fixed neighborhood range. Multi-scale embedding encodes nearby edges, small object textures, and larger neighborhood context before frequency-domain aggregation. IP and YZ obtain larger gains because crop parcels and large-scene objects appear at different spatial scales.

4.3.3. Comparison of Frequency-Domain Feature Extraction Strategies in HSFIM

Fourier filtering is crucial for HSFIM, as it simultaneously enables global feature aggregation and extracts the amplitude and phase information required by the subsequent FFAB. To verify its effectiveness, we compared our proposed learnable complex Fourier filtering (FFT) with four alternative strategies: (1) Spatial Conv, which retains only the 3×3 spatial convolution; (2) Large-kernel Conv, which uses a 5×5 convolution; (3) DCT, which applies learnable filtering to discrete cosine coefficients; and (4) DWT, which performs wavelet decomposition, learnable subband modulation, and inverse reconstruction. To ensure a fair comparison, only the feature extraction branch was modified while all other network components and training configurations remained strictly identical, with the quantitative results summarized in Table 11.

Table 11. Comparison of feature extraction strategies in HSFIM in terms of OA.

Feature Extractor	IP OA (%)	SA OA (%)	PU OA (%)	YZ OA (%)
Spatial Conv	52.32	75.43	49.72	46.54
Large-kernel Conv	51.60	77.42	50.79	53.32
DCT	52.50	75.01	53.62	59.27
DWT	52.88	74.96	47.86	58.97
FFT	57.47	79.38	54.57	64.11

Note: Bold values indicate the highest OA on each dataset.

As shown in Table 11, FFT achieves the highest OA on all four datasets. Compared with Spatial Conv, Large-kernel Conv, DCT, and DWT, the OA gains of FFT range from 4.59% to 5.87% on Indian Pines, from 1.96% to 4.42% on Salinas, from 0.95% to 6.71% on Pavia University, and from 4.84% to 17.57% on Yangzhou. On the large-scale Yangzhou dataset, FFT exceeds Spatial Conv, Large-kernel Conv, DCT, and DWT by 17.57%, 10.79%, 4.84%, and 5.14%, respectively. The quantitative comparison shows that effective feature extraction in HSFIM requires both full-image context modeling and amplitude–phase representation. Spatial Conv and Large-kernel Conv mainly capture local spatial patterns and focus primarily on short-range regional relationships. DCT represents global frequency information through cosine bases. FFT further preserves both amplitude and phase, matching the amplitude recalibration and phase preservation operations in FFAB. DWT focuses on localized multi-scale subbands and provides detailed local frequency features. By combining full-image frequency aggregation with amplitude–phase preservation, FFT achieves higher OA than the four alternative feature extractors on all datasets.

4.3.4. Phase Sensitivity Analysis of FreqViT

The query–key cross-spectrum in FreqViT contains magnitude and phase responses. Phase-related variants are constructed to examine whether explicit phase modulation is necessary for the attention mask. Given the query–key cross-spectrum

$$\mathbf{R}_f = \mathbf{Q}_f \odot \text{conj}(\mathbf{K}_f), \quad (20)$$

the magnitude response and unit phase response are defined as

$$\mathbf{M}_f = |\mathbf{R}_f|, \quad \mathbf{U}_f = \frac{\mathbf{R}_f}{|\mathbf{R}_f| + \epsilon}, \quad (21)$$

with ϵ representing a small constant for numerical stability. Attention masks for frequency-domain variants are defined as

$$\mathbf{A}_f^{\text{mag}} = \text{Softmax}_f(\mathbf{M}_f), \quad (22)$$

$$\mathbf{A}_f^{\text{real}} = \text{Softmax}_f(\text{Re}(\mathbf{R}_f)), \quad (23)$$

$$\mathbf{A}_f^{\text{detach}} = \text{Softmax}_f(\mathbf{M}_f) \odot \text{sg}(\mathbf{U}_f), \quad (24)$$

$$\mathbf{A}_f^{\text{grad}} = \text{Softmax}_f(\mathbf{M}_f) \odot \mathbf{U}_f, \quad (25)$$

$$\mathbf{A}_f^{\text{reverse}} = \text{Softmax}_f(\mathbf{M}_f) \odot \text{Flip}_f(\text{sg}(\mathbf{U}_f)), \quad (26)$$

$$\mathbf{A}_f^{\text{random}} = \text{Softmax}_f(\mathbf{M}_f) \odot e^{j\theta}, \quad \theta \sim \mathcal{U}(-\pi, \pi), \quad (27)$$

$$\mathbf{A}_f^{\text{phase}} = \mathbf{U}_f. \quad (28)$$

Here, $\text{sg}(\cdot)$ denotes stop-gradient, and $\text{Flip}_f(\cdot)$ denotes reversal along frequency dimension. Spatial-attention baseline is denoted as *Spatial-Base*, and attention mask is defined as

$$\mathbf{A}^{\text{spa}} = \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_h}). \quad (29)$$

Table 12 reports OA for all variants.

Table 12. OA comparison of phase-related FreqViT variants in the SFI-APD framework (%).

Method	IP OA (%)	SA OA (%)	PU OA (%)	YZ OA (%)
<i>Spatial-Base</i>	49.73	71.25	51.85	51.72
<i>Freq-Mag</i>	57.47	79.38	54.57	64.11
<i>Freq-Real</i>	47.90	75.53	53.22	64.43
<i>Freq-Phase-SG</i>	51.31	72.97	51.35	64.32
<i>Freq-Phase-Grad</i>	47.56	76.28	41.33	53.29
<i>Freq-Phase-Reverse</i>	51.34	75.85	47.96	52.98
<i>Freq-Phase-Random</i>	50.45	74.18	48.50	47.58
<i>Freq-Phase-Only</i>	40.50	57.16	50.90	51.58

Note: Bold values indicate the highest OA on each dataset.

Table 12 indicates *Freq-Mag* obtains highest OA on IP, SA, and PU, with 57.47%, 79.38%, and 54.57%, respectively. On YZ, *Freq-Real* obtains 64.43% OA, and *Freq-Phase-SG* obtains 64.32% OA; margins over *Freq-Mag* are 0.32% and 0.21%. Direct phase-gradient modulation reduces OA to 47.56% on IP, 41.33% on PU, and 53.29% on YZ. Random-phase and phase-only variants also give unstable results. FreqViT therefore uses magnitude-based cross-spectrum attention for token aggregation, and FFAB preserves phase information during reconstruction for boundary geometry.

4.4. Attention Mechanism Analysis

Tables 13–16 compare FreqViT with Spatial Attention, Performer [14], CosFormer [36], AFT [37], FNet [38], GFNet [39], DCT [40], and FFTformer-FSAS [41] under the same backbone, hyperparameters, and training strategy. In complexity expressions, N , D , r , and K denote token number, token dimension, Performer random-feature number, and DCT retained frequency-component number, respectively.

Table 13 reports FreqViT obtains 57.47% OA, 57.88% NMI, 39.71% ARI, and 46.83% FMI on IP. Performer gives 55.78% OA, and FNet/GFNet give 51.81% OA with lower FLOPs. Frequency mixing reduces computation, and query–key cross-spectrum modulation improves cluster assignment and pairwise grouping for mixed crop parcels.

Table 14 reports FreqViT obtains 79.38% OA, 90.37% NMI, 72.99% ARI, and 78.33% FMI on SA. FNet/GFNet reach 77.42% OA with fewer parameters, but NMI, ARI, and FMI remain lower than FreqViT. Cross-spectrum modulation improves partition agreement in regular crop parcels with high spectral continuity.

Table 13. Comparison of different attention mechanisms on Indian Pines.

Method	Complexity	OA (%)	NMI (%)	ARI (%)	AMI (%)	FMI (%)	FLOPs	Params
FreqViT	$\mathcal{O}(ND \log N)$	57.47	57.88	39.71	57.68	46.83	76.497G	203.301K
Spatial Attention	$\mathcal{O}(N^2D)$	49.74	55.97	32.99	55.76	40.97	79.114G	203.301K
FNet	$\mathcal{O}(ND \log N)$	51.81	54.10	33.03	53.90	41.15	56.904G	165.861K
Performer	$\mathcal{O}(NrD)$	55.78	53.61	37.67	53.39	46.24	81.046G	203.301K
CosFormer	$\mathcal{O}(ND^2)$	50.15	56.56	32.84	56.35	40.58	77.578G	203.301K
GFNet	$\mathcal{O}(ND \log N)$	51.81	54.12	33.03	53.90	41.08	56.904G	165.861K
FFTformer-FSAS	$\mathcal{O}(ND \log N)$	52.42	56.78	34.93	56.57	42.94	76.881G	203.685K
AFT	$\mathcal{O}(N^2D)$	51.03	56.04	33.90	55.83	41.58	79.114G	215.589K
DCT	$\mathcal{O}(ND \log N + KD)$	50.50	56.04	33.53	55.83	41.16	61.734G	203.301K

Note: Bold values indicate the best result in each column. Lower values are better for FLOPs and Params, whereas higher values are better for the clustering metrics.

Table 14. Comparison of different attention mechanisms on Salinas.

Method	Complexity	OA (%)	NMI (%)	ARI (%)	AMI (%)	FMI (%)	FLOPs	Params
FreqViT	$\mathcal{O}(ND \log N)$	79.38	90.37	72.99	90.37	78.33	76.497G	203.301K
Spatial Attention	$\mathcal{O}(N^2D)$	74.55	86.95	69.28	86.94	75.51	79.114G	203.301K
FNet	$\mathcal{O}(ND \log N)$	77.42	86.85	70.41	86.83	75.98	56.904G	165.861K
Performer	$\mathcal{O}(NrD)$	77.01	86.58	70.09	86.57	75.74	81.046G	203.301K
CosFormer	$\mathcal{O}(ND^2)$	72.79	85.42	67.40	85.40	73.20	77.578G	203.301K
GFNet	$\mathcal{O}(ND \log N)$	77.42	86.84	70.41	86.83	75.97	56.904G	165.861K
FFTformer-FSAS	$\mathcal{O}(ND \log N)$	76.16	87.06	70.65	87.05	76.52	76.881G	203.685K
AFT	$\mathcal{O}(N^2D)$	74.29	87.22	69.37	87.21	75.59	79.114G	215.589K
DCT	$\mathcal{O}(ND \log N + KD)$	74.22	87.52	69.81	87.51	75.80	61.734G	203.301K

Note: Bold values indicate the best result in each column. Lower values are better for FLOPs and Params, whereas higher values are better for the clustering metrics.

Table 15. Comparison of different attention mechanisms on Pavia University.

Method	Complexity	OA (%)	NMI (%)	ARI (%)	AMI (%)	FMI (%)	FLOPs	Params
FreqViT	$\mathcal{O}(ND \log N)$	54.57	53.28	39.40	53.26	52.38	76.497G	203.301K
Spatial Attention	$\mathcal{O}(N^2D)$	48.95	53.70	34.45	53.68	48.33	79.114G	203.301K
FNet	$\mathcal{O}(ND \log N)$	51.54	51.56	36.27	51.54	49.80	56.904G	165.861K
Performer	$\mathcal{O}(NrD)$	49.18	53.51	33.65	53.49	47.69	81.046G	203.301K
CosFormer	$\mathcal{O}(ND^2)$	49.87	54.36	34.77	54.34	48.60	77.578G	203.301K
GFNet	$\mathcal{O}(ND \log N)$	52.08	51.56	36.78	51.54	50.23	56.904G	165.861K
FFTformer-FSAS	$\mathcal{O}(ND \log N)$	52.95	52.26	37.27	52.24	50.56	76.881G	203.685K
AFT	$\mathcal{O}(N^2D)$	51.08	55.12	35.97	55.10	49.58	79.114G	215.589K
DCT	$\mathcal{O}(ND \log N + KD)$	52.47	53.02	38.44	53.00	51.59	61.734G	203.301K

Note: Bold values indicate the best result in each column. Lower values are better for FLOPs and Params, whereas higher values are better for the clustering metrics.

Table 15 reports FreqViT reaches 54.57% OA, 39.40% ARI, and 52.38% FMI on PU. FFTformer-FSAS gives 52.95% OA, and GFNet gives 52.08% OA with fewer parameters. AFT obtains higher NMI and AMI, but OA and ARI remain 51.08% and 35.97%. Frequency query–key modulation gives higher assignment accuracy and pairwise consistency for urban material clustering.

Table 16 reports FreqViT obtains 64.11% OA, 53.37% NMI, 46.00% ARI, and 57.69% FMI on YZ. CosFormer and DCT reach 60.11% and 58.25% OA, respectively. FFTformer-FSAS gives 52.20% NMI and 42.60% ARI. FNet/GFNet reduce parameter count to 165.861K, with 57.81% OA. For large HSI scenes, frequency query–key cross-correlation provides higher long-range aggregation accuracy than plain spectral filtering, linearized attention, or retained-component transforms.

Across four datasets, FreqViT gives higher OA and partition-consistency metrics in most cases. Spatial Attention preserves pairwise token relations but requires quadratic affinity computation. FNet, GFNet, and DCT reduce cost through Fourier mixing or component retention, but query–key dependency becomes weaker. Performer and CosFormer reduce attention cost through approximation or linearization, which can affect mixed pixels

around boundaries. FreqViT keeps query–key relation by cross-spectrum modulation without storing an $N \times N$ affinity matrix, forming an accuracy–cost balance for SFI-APD.

Table 16. Comparison of different attention mechanisms on Yangzhou.

Method	Complexity	OA (%)	NMI (%)	ARI (%)	AMI (%)	FMI (%)	FLOPs	Params
FreqViT	$\mathcal{O}(ND \log N)$	64.11	53.37	46.00	53.37	57.69	76.497G	203.301K
Spatial Attention	$\mathcal{O}(N^2 D)$	49.72	39.30	29.82	39.29	45.79	79.114G	203.301K
FNet	$\mathcal{O}(ND \log N)$	57.81	49.81	40.85	49.81	53.20	56.904G	165.861K
Performer	$\mathcal{O}(NrD)$	48.22	38.16	28.27	38.16	44.79	81.046G	203.301K
CosFormer	$\mathcal{O}(ND^2)$	60.11	48.33	38.44	48.33	52.00	77.578G	203.301K
GFNet	$\mathcal{O}(ND \log N)$	57.81	49.81	40.85	49.81	53.20	56.904G	165.861K
FFTformer-FSAS	$\mathcal{O}(ND \log N)$	59.11	52.20	42.60	52.19	54.64	76.881G	203.685K
AFT	$\mathcal{O}(N^2 D)$	52.07	42.15	30.83	42.15	46.48	79.114G	215.589K
DCT	$\mathcal{O}(ND \log N + KD)$	58.25	51.26	42.46	51.26	54.54	61.734G	203.301K

Note: Bold values indicate the best result in each column. Lower values are better for FLOPs and Params, whereas higher values are better for the clustering metrics.

4.5. Sensitivity Analysis of PCA Dimension

PCA is utilized to reduce spectral redundancy and extract the most informative spectral features prior to training the proposed network. Since the number of retained principal components significantly impacts the preservation of spectral information, we conducted a systematic sensitivity analysis by varying the PCA dimension $d \in \{10, 20, 30, 40, 50\}$ across the four datasets.

As observed in Figure 9, increasing the PCA dimension from 10 to 30 leads to a general improvement in unsupervised clustering accuracy across the datasets, as a higher dimension preserves more discriminative spectral information. However, when d exceeds 30, the performance tends to stabilize or decrease. Although a larger dimension includes more spectral information, it inevitably introduces redundant spectral components and noise associated with higher-order principal components, which can interfere with the unsupervised clustering process. Specifically, as shown in Figure 9, the proposed network achieves the highest clustering accuracy (OA) at $d = 30$ for the IP and PU datasets. For the SA dataset, the accuracy peaks slightly higher at $d = 40$, whereas for the YZ dataset, the optimal dimension is $d = 20$. Despite these minor dataset-specific variations, setting $d = 30$ yields consistently competitive and robust performance across all four datasets. Therefore, to achieve an optimal balance among the preservation of essential spectral information, the suppression of noise, computational efficiency, and experimental consistency, we selected $d = 30$ as the unified PCA dimension for all experiments.

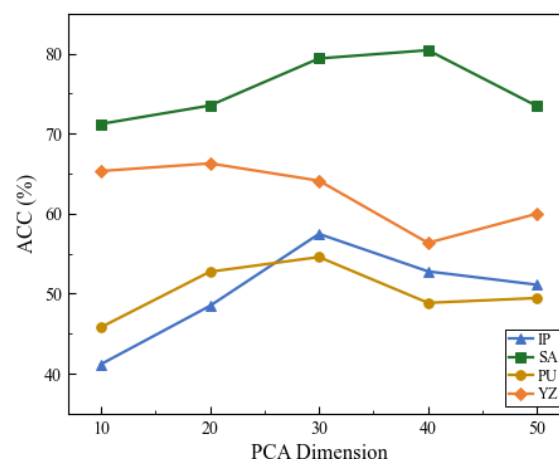


Figure 9. Sensitivity analysis of the PCA dimension on the four datasets in terms of clustering accuracy (OA).

5. Discussion

Experimental results across four HSI datasets show consistent benefits from spatial–frequency interaction and amplitude–phase decoupling. Gains on IP and YZ are larger than gains on SA and PU. IP contains mixed crop parcels and small categories, and YZ contains large spatial extent with roads, buildings, tidal flats, vegetation, and water-related surfaces. HSFIM supplies local boundary responses through convolution and full-image aggregation through Fourier filtering, so separated regions with similar spectra can be linked without dense token-pair affinity.

Ablation results in Tables 9–12 explain module behavior. Removing HSFIM reduces OA on four datasets because local convolution alone cannot provide global frequency context. Removing FFAB reduces accuracy because illumination-sensitive amplitude changes remain coupled with boundary-related phase information. Replacing FreqViT with Spatial Attention also reduces accuracy, suggesting frequency-domain query–key modulation contributes to long-range aggregation rather than serving only as a low-cost substitute for standard attention.

Attention comparisons in Tables 13–16 show an accuracy–cost pattern. FNet, GFNet, and DCT reduce computation through frequency mixing or retained frequency components, but query–key dependency becomes weaker. Spatial Attention keeps dense pairwise relations and needs quadratic affinity storage. FreqViT keeps query–key correlation through cross-spectrum modulation and avoids an $N \times N$ attention matrix, which is suitable for large HSI scenes.

6. Conclusions

SFI-APD presents a self-supervised framework for clustering-friendly feature learning in HSI clustering. By integrating HSFIM, FFAB, and FreqViT, SFI-APD captures local spatial details, global frequency dependencies, and illumination-robust spectral–spatial representations within a unified architecture. HSFIM combines convolutional texture extraction with learnable Fourier filtering to enhance local–global feature interaction. FFAB decouples amplitude and phase components, recalibrates illumination-sensitive amplitude responses, and preserves phase-related boundary geometry. FreqViT performs query–key cross-spectrum modulation in frequency domain, reducing token aggregation complexity from $\mathcal{O}(N^2D)$ to $\mathcal{O}(ND \log N)$ and improving long-range contextual modeling. Experimental results on Indian Pines, Salinas, Pavia University, and Yangzhou datasets demonstrate SFI-APD improves clustering accuracy and spatial region continuity compared with representative HSI clustering methods.

Although SFI-APD has achieved excellent clustering performance, certain limitations remain in complex terrain analysis. During the iterative optimization of clustering centers, SFI-APD shows limited capability in identifying minority classes, small-scale targets, and highly similar spectral features. Misclassification often occurs in areas with strong spectral similarity, such as alfalfa and oats in the Indian Pines dataset, as well as lettuce at different growth stages in the Salinas dataset. Meanwhile, complex scene spatial topology reduces the clarity of local clustering boundaries. This issue is especially evident in the shadow areas of the Pavia University dataset and the water–land junction areas of the Yangzhou dataset.

To address these issues, future work will explore adaptive cluster number estimation to reduce dependence on prior hyperparameters. Boundary-aware self-supervised constraints will also be introduced to better separate fine-grained neighboring regions. In addition, multi-source remote sensing data, such as LiDAR, will be used for cross-modal interaction to enhance feature edges between spectrally similar materials and improve overall clustering robustness.

Author Contributions: Conceptualization, H.Y., N.H. and Z.Z.; methodology, H.Y.; software, H.Y.; validation, H.Y., N.H., Q.L., P.L., K.N. and Z.Z.; formal analysis, H.Y.; investigation, H.Y.; resources, N.H. and Z.Z.; data curation, H.Y. and Q.L.; writing—original draft preparation, H.Y.; writing—review and editing, N.H., Q.L., P.L., K.N. and Z.Z.; visualization, H.Y.; supervision, N.H. and Z.Z.; project administration, N.H. and Z.Z.; funding acquisition, N.H. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Research was funded in part by the National Natural Science Foundation of China under Grant 62501296, 62501308, and 62471235; in part by the Frontier Technologies R&D Program of Jiangsu under Grant BF2024070; in part by the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant 23KJB510022 and 23KJB520026; in part by Deep Earth Probe and Mineral Resources Exploration-National Science and Technology Major Project under Grant 2025ZD1008400; in part by the Jiangsu Geological Bureau Research Project under Grant 2023KY11; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20230914; and in part by the Jiangsu Association for Science and Technology Youth Talent Promotion Program under Grant JSTJ2025138.

Data Availability Statement: Public Indian Pines, Salinas, and Pavia University hyperspectral datasets used in the study are available from public repositories. The Yangzhou dataset is available from the corresponding author upon reasonable request, subject to the permission of the data provider.

Acknowledgments: The authors thank the Engineering Research Center in Jiangsu Province for providing the high-resolution Yangzhou hyperspectral dataset. The dataset was acquired using the airborne Multi-Parameter Airborne Remote Sensing System (MARSS) and supported large-scale evaluation in the study. The authors are grateful to anonymous reviewers and the Associate Editor for constructive comments and suggestions. Reviewer feedback improved presentation and technical clarity of the manuscript. The authors also thank researchers who released source code of comparison methods, which supported comprehensive experimental evaluation.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. <https://doi.org/10.1109/MGRS.2017.2762087>.
2. Luo, F.; Hua, Y.; Fu, C.; Guo, T.; Duan, Y.; Shi, G. MMD-MLP: LiDAR-guided hyperspectral data classification using local-global directional-MLP with multi-resolution multi-scale representation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14. <https://doi.org/10.1109/TGRS.2025.3550370>.
3. Zhou, T.; Luo, F.; Fu, C.; Guo, T.; Du, B.; Gao, X.; Zhang, L. TSCCD: Temporal self-construction cross-domain learning for unsupervised hyperspectral change detection. *IEEE Trans. Image Process.* **2026**, *35*, 830–843. <https://doi.org/10.1109/TIP.2025.3650387>.
4. Luo, F.; Chen, X.; Fu, C.; Guo, T.; Du, B. HDiff-HIR: Hierarchically conditional diffusion model for hyperspectral image reconstruction. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 777–791. <https://doi.org/10.1109/TCSVT.2025.3597548>.
5. Luo, F.; Chen, X.; Guo, T.; Gong, X.; Zhang, L.; Zhu, C. DSMT: Dual-stage multiscale Transformer for hyperspectral snapshot compressive imaging. *IEEE Trans. Image Process.* **2025**, *34*, 2473–2486. <https://doi.org/10.1109/TIP.2025.3556520>.
6. Lei, J.; Li, X.; Peng, B.; Fang, L.; Ling, N.; Huang, Q. Deep spatial-spectral subspace clustering for hyperspectral image. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2686–2697. <https://doi.org/10.1109/TCSVT.2020.3027616>.
7. Hu, X.; Li, T.; Zhou, T.; Peng, Y. Deep spatial-spectral subspace clustering for hyperspectral images based on contrastive learning. *Remote Sens.* **2021**, *13*, 4418. <https://doi.org/10.3390/rs13214418>.
8. Liu, B.; Kong, W.; Wang, Y. Deep convolutional asymmetric autoencoder-based spatial-spectral clustering network for hyperspectral image. *Wireless Commun. Mobile Comput.* **2022**, *2022*, 1–12. <https://doi.org/10.1155/2022/2027981>.
9. Zheng, Z.; Zhang, S.; Song, H.; Yan, Q. Deep clustering using 3D attention convolutional autoencoder for hyperspectral image analysis. *Sci. Rep.* **2024**, *14*, 4209. <https://doi.org/10.1038/s41598-024-54547-2>.

10. Luo, F.; Liu, Y.; Duan, Y.; Guo, T.; Zhang, L.; Du, B. SDST: Self-supervised double-structure Transformer for hyperspectral images clustering. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. <https://doi.org/10.1109/TGRS.2024.3374597>.
11. Huang, N.; Li, C.; Liu, Q.; Li, J.; Liu, P.; Zheng, Z.; Xiao, L. A semantic Transformer for self-supervised deep clustering of large hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5512017. <https://doi.org/10.1109/TGRS.2026.3686883>.
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>.
13. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A general vision Transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
14. Choromanski, K.; Likhoshesterov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; Belanger, D.; Colwell, L.; Weller, A. Rethinking attention with Performers. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
15. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5156–5165.
16. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Adaptive Fourier neural operators: Efficient token mixers for Transformers. In Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022.
17. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; Mei, T. Wave-ViT: Unifying wavelet and Transformers for visual representation learning. In *Computer Vision—ECCV 2022, Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2022; Volume 13685, pp. 328–345. https://doi.org/10.1007/978-3-031-19806-9_19.
18. Patro, B.N.; Namboodiri, V.P.; Agneeswaran, V.S. SpectFormer: Frequency and attention is what you need in a Vision Transformer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 26 February–6 March 2025; pp. 9543–9554. <https://doi.org/10.1109/WACV61041.2025.00924>.
19. Wang, C.; Wu, H.; Jin, Z. FourLLIE: Boosting low-light image enhancement by Fourier frequency information. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, Canada, 29 October–3 November 2023; pp. 7459–7469. <https://doi.org/10.1145/3581783.3611909>.
20. Oppenheim, A.V.; Lim, J.S. The importance of phase in signals. *Proc. IEEE* **1981**, *69*, 529–541. <https://doi.org/10.1109/PROC.1981.12022>.
21. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York City, NY, USA, 19–24 June 2016; pp. 478–487.
22. Guo, X.; Gao, L.; Liu, X.; Yin, J. Improved deep embedded clustering with local structure preservation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017), Melbourne, Australia, 19–25 August 2017; pp. 1753–1759. <https://doi.org/10.24963/ijcai.2017/243>.
23. Yang, B.; Fu, X.; Sidiropoulos, N.D.; Hong, M. Towards K-means-friendly spaces: Simultaneous deep learning and clustering. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017; pp. 3861–3870.
24. Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; Zhou, H. Variational deep embedding: An unsupervised and generative approach to clustering. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), Melbourne, Australia, 19–25 August 2017; pp. 1965–1972. <https://doi.org/10.24963/ijcai.2017/273>.
25. Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; Reid, I. Deep subspace clustering networks. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 24–33.
26. Cai, J.; Fan, J.; Guo, W.; Wang, S.; Zhang, Y.; Zhang, Z. Efficient deep embedded subspace clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 18–24 June 2022; pp. 21–30. <https://doi.org/10.1109/CVPR52688.2022.00012>.
27. Wu, W.; Wang, W.; Kong, S. Deep structure and attention aware subspace clustering. In Proceedings of the 6th Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2023), Xiamen, China, 13–15 October 2023; pp. 139–150. https://doi.org/10.1007/978-981-99-8462-6_12.
28. Zhang, Y.; Wang, Y.; Chen, X.; Jiang, X.; Zhou, Y. Spectral-spatial feature extraction with dual graph autoencoder for hyperspectral image clustering. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8500–8511. <https://doi.org/10.1109/TCSVT.2022.3196679>.
29. Xu, Y.; Huang, D.; Wang, C.-D.; Lai, J.-H. Deep image clustering with contrastive learning and multi-scale graph convolutional networks. *Pattern Recognit.* **2024**, *146*, 110065. <https://doi.org/10.1016/j.patcog.2023.110065>.
30. Wang, J.; Guan, R.; Gao, K.; Li, Z.; Li, H.; Li, X.; Tang, C. Multi-level graph subspace contrastive learning for hyperspectral image clustering. In Proceedings of the International Joint Conference on Neural Networks (IJCNN 2024), Yokohama, Japan, 30 June–5 July 2024; pp. 1–8. <https://doi.org/10.1109/IJCNN60899.2024.10650148>.

31. Cai, Y.; Liu, Y.; Zhang, Z.; Cai, Z.; Liu, X. Large-scale hyperspectral image clustering using contrastive learning. In Proceedings of the 1st Workshop on Complex Data Challenges in Earth Observation (CDCEO 2021), Virtual, 1 November 2021; Volume 3052.
32. Li, J.; Guan, R.; Han, Y.; Hu, Y.; Li, Z.; Wu, Y.; Xu, Z.; Li, X. Superpixel-based dual-neighborhood contrastive graph autoencoder for deep subspace clustering of hyperspectral image. In *Advanced Intelligent Computing Technology and Applications, Lecture Notes in Computer Science*, 14867; Springer: Singapore, 2024; pp. 181–192. https://doi.org/10.1007/978-981-97-5597-4_16.
33. Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; Feng, R. Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. <https://doi.org/10.1109/TGRS.2024.3370633>.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
36. Qin, Z.; Sun, W.; Deng, H.; Li, D.; Wei, Y.; Lv, B.; Yan, J.; Kong, L.; Zhong, Y. cosFormer: Rethinking softmax in attention. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022.
37. Zhai, S.; Talbott, W.; Srivastava, N.; Huang, C.; Goh, H.; Zhang, R.; Susskind, J. An attention free Transformer. *arXiv* **2021**, arXiv:2105.14103. <https://doi.org/10.48550/arXiv.2105.14103>.
38. Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; Ontanon, S. FNet: Mixing tokens with Fourier transforms. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2022), Seattle, WA, USA, 10–15 July 2022; pp. 4296–4313. <https://doi.org/10.18653/v1/2022.naacl-main.319>.
39. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 980–993.
40. Pan, H.; Hamdan, E.; Zhu, X.; Biswas, K.; Cetin, A.E.; Bagci, U. DCT-based decorrelated attention for vision Transformers. *arXiv* **2024**, arXiv:2405.13901. <https://doi.org/10.48550/arXiv.2405.13901>.
41. Kong, L.; Dong, J.; Ge, J.; Li, M.; Pan, J. Efficient frequency domain-based Transformers for high-quality image deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, BC, Canada, 17–24 June 2023; pp. 5886–5895.
42. Tadjudin, S.; Landgrebe, D.A. Covariance estimation with limited training samples. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 2113–2118. <https://doi.org/10.1109/36.774728>.
43. Vane, G.; Green, R.O.; Chrien, T.G.; Enmark, H.T.; Hansen, E.G.; Porter, W.M. The airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sens. Environ.* **1993**, *44*, 127–143. [https://doi.org/10.1016/0034-4257\(93\)90012-M](https://doi.org/10.1016/0034-4257(93)90012-M).
44. Palmason, J.A.; Benediktsson, J.A.; Sveinsson, J.R. Classification of hyperspectral ROSIS data from urban areas. In Proceedings of the 2nd International Conference on Recent Advances in Space Technologies (RAST 2005), Istanbul, Turkey, 9–11 June 2005; pp. 63–69. <https://doi.org/10.1109/RAST.2005.1512536>.
45. Huang, N.; Wu, T.; Liu, Q.; Liu, P.; Ni, K.; Zheng, Z.; Xiao, L. CNLTrans: A Clusterwise Nonlocal Transformer for Self-Supervised Deep Clustering of Large Hyperspectral Images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2026**, *19*, 21295–21311. <https://doi.org/10.1109/JSTARS.2026.3702236>.
46. Guo, W.; Lin, K.; Ye, W. Deep embedded K-means clustering. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW 2021), Auckland, New Zealand, 7–10 December 2021; pp. 686–694. <https://doi.org/10.1109/ICDMW53433.2021.00090>.
47. Yang, X.; Liu, Y.; Zhou, S.; Wang, S.; Tu, W.; Zheng, Q.; Liu, X.; Fang, L.; Zhu, E. Cluster-guided contrastive graph clustering network. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 10834–10842. <https://doi.org/10.1609/aaai.v37i9.26285>.
48. Ding, Y.; Zhang, Z.; Yang, A.; Cai, Y.; Xiao, X.; Hong, D.; Yuan, J. SLCGC: A lightweight self-supervised low-pass contrastive graph clustering network for hyperspectral images. *IEEE Trans. Multimedia* **2025**, *27*, 8251–8262. <https://doi.org/10.1109/TMM.2025.3604954>.
49. Qian, Q. Stable cluster discrimination for deep clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 16599–16608. <https://doi.org/10.1109/ICCV51070.2023.01526>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.